Evaluation in Polycentric Governance Systems: Climate Change Policy in the European Union

Jonas J. Schoenefeld, BA, MPhil

A thesis submitted to the School of Environmental Sciences of the University of East Anglia in partial fulfilment of the requirements for the Degree of Doctor of Philosophy.

April 2018

For my parents, Rosemarie Schönefeld and Stephan Schönefeld-Dorka, and for my partner, Nicole Validiva Rebolledo, without whose love and support none of this would have been possible.

Abstract

Perceived failures in top-down climate governance and many emerging bottom-up activities have prompted scholars to pay more attention to the promise and limits of polycentric governance, in which activities are spread across many levels, actors, and scales (E. Ostrom, 2010c; E. Ostrom, 2014b). In adopting the Paris Agreement, policy makers also appear to be moving in the direction of greater polycentricity. But many aspects of polycentric governance remain theoretically and empirically underexplored, especially with a view to policy evaluation, a vital but often neglected governance activity. This thesis addresses these gaps by: (1) considering the potential (theoretical) role of policy evaluation in polycentric governance and (2) empirically exploring the case of the European Union, an active adopter and evaluator of climate policy whose climate governance has been described as polycentric. The thesis argues that polycentric governance theory is based on three foundational ideas, namely that that actors can and do self-organize, that context matters in governance, and that governance centres, while independent, interact in order to fully realize the benefits of polycentric governance. These foundational ideas provide a means to explore climate policy evaluation, and to connect with related debates in the evaluation literatures. Fresh empirical data from a new database of 618 climate policy evaluations (1997-2014) suggest that formal (state) actors produced many more evaluations that informal (societal) ones—pointing to limited self-organization and a key role for public actors in evaluation—but that informal evaluations also emerged in empirically detectable and relevant quantities. By using a new coding scheme to analyse a sub-set of the evaluations this thesis reveals that the limited attention to various contextual factors and the fact that climate policy evaluation tends to happen in and focus on individual governance centres restricts the potential travel of evaluative insights from one governance centre to another. *In toto*, the empirical characteristics equip climate policy evaluation only partially to facilitate polycentric climate governance in the EU.

Contents

Acknowledgements

As I have come to learn, writing a PhD thesis is never a single project or, for that matter, a straight line. Rather, it is a jigsaw or a mosaic of many different pieces of insight and learning that ultimately find their way into the final thesis. There are many, many people who have generously offered helpful thoughts, feedback, advice and, of course, encouragement, to keep going with what seemed like a monumental task at times.

When navigating unfamiliar territory, there is nothing more useful than a knowledgeable and experienced guide. First and foremost, I would like to thank my supervisor Professor Andrew Jordan, who has not only worked tirelessly to support my PhD process, but who has also opened up numerous other opportunities that have significantly contributed to this project. Working with Andy has been immensely intellectually stimulating and enjoyable, and I thank him for his support, trust in my abilities, and also his contestation.

As my second supervisor, Dr Irene Lorenzoni has been equally supportive and she has offered critical advice and ideas, support and also understanding at countless points over the years in which I wrote this thesis. Thank you, Irene.

I would furthermore like to thank the members of the Innovations in Climate Governance Network (INOGOV), which proved a wonderful way to meet many colleagues, and contributed significantly to the various publication projects that flank my PhD (Jordan et al., 2015; Jordan, Huitema, Schoenefeld, van Asselt, & Forster, 2018; Schoenefeld & Jordan, 2017; Schoenefeld, Hildén, & Jordan, 2018; Schoenefeld & McCauley, 2016; Tosun & Schoenefeld, 2017). These publications created valuable synergies with my PhD by allowing me to explore aspects that would have gone well beyond the reach of this thesis, but were nevertheless highly relevant in scoping the field and tapping new venues for productive research. As I was navigating vast and at least to

me in part unknown literatures, I was blessed with many constructive and patient peer reviewers, who saw a valuable contribution in my work and were willing to help me address gaps in my knowledge and understanding. It was also very instructive to conduct various publication projects with more experienced, senior scholars, most notably Professor Andrew Jordan, but also Professor Mikael Hildén, Professor Jale Tosun, Professor Michelle McCauley, and others. I also had the privilege of coordinating INOGOV's Early Career Investigators' Network (ECIN), and thank its members for their enthusiasm and collaboration.

I am grateful to Professor Dave Huitema and Professor Mikael Hildén for offering early advice on my coding scheme. Christoph Priebe generously agreed to be a second coder on a number of evaluations in order to test the inter-coder reliability of my coding scheme and calibrate it in the process. Mikael hosted me for a three-month research fellowship at the Finnish Environment Institute in spring 2015, where we explored other aspects of policy monitoring and evaluation, and ultimately wrote an article for *Climate Policy* (Schoenefeld et al., 2018). Thomas Delahais and Dr Stijn van Voorst deserve credit for respectively allowing me to access and use the 'Eureval' and the '*Ex-post* Legislative Evaluations' evaluation databases and Dr Stefania Sitzia for making helpful suggestions on sampling techniques.

Institutionally, the Tyndall Centre for Climate Change Research and the School of Environmental Sciences at the University of East Anglia proved a highly intellectually stimulating, welcoming, and supportive environment. Being a student here became a lived experience of the notion that, as Thomas Piketty (2014, p. 32-33) ably put it, "[t]he social sciences collectively know too little to waste time on foolish disciplinary squabbles." I very much thrived in an environment that treasured interdisciplinary exchange, much of which happens in the corridors, over lunch or—in the fine British tradition—in the pub. I learned a great deal about the broader facets of climate change, even though being at an interdisciplinary school also required regular conference travel in order to exchange with colleagues in one's own highly specialized field.

I am grateful to the many other Tyndall colleagues who made my time in Norwich so enjoyable, with special thanks to Adam Kennerley for many walks and talks, and to

Dr Viviane Gravey and Brendan Moore for collaborating so well on the Environmental Europe? Blog.[1] Tyndall Director Professor Corinne Le Quéré and Executive Director Asher Minns deserve special credit for being open to new ideas and supporting various initiatives to consider the sustainability of our own operations. I also thank Dr Peter Simmons for his kind support in his role as Director of Graduate Research at the School of Environmental Sciences.

I thank the participants of the conferences and workshops, which I attend while writing this thesis:

- UACES Student Forum in Birmingham, UK, in April 2014
- Global Multi-Level Climate Governance Conference at the IASS in Potsdam, Germany, in September 2014
- International Energy Policy & Programme Evaluation Conference in Berlin, Germany, in September 2014
- UACES Student Forum, London, UK, in November 2014
- INOGOV Governance Experiments Workshop, Helsinki, Finland, in March 2015
- UACES General Conference in Bilbao, Spain, in September 2015
- European Environmental Evaluators' Network (EEEN) Forum in Florence, Italy, in September 2015
- INOGOV ECIN Workshop on 'Data Frontiers in Climate Governance Research' in Zurich, Switzerland, in February 2016
- ECPR Workshop on Environmental Policy in the EU in Pisa, Italy, in April 2016
- ECPR Standing Group Conference on Regulatory Governance in Tilburg, the Netherlands, in July 2016
- MMR Workshop at the European Environment Agency in Copenhagen, Denmark, in September 2016
- UACES Annual Conference in Krakow, Poland, September 2017
- ECPR General Conference in Oslo, Norway, September 2017

[1] http://environmentaleurope.ideasoneurope.eu/

- Workshop on Policy Evaluation in the EU in Brussels, Belgium, January 2018
- INOGOV Final Action Conference, Brussels, Belgium, in March 2018

I furthermore thank the School of Environmental Sciences at the University of East Anglia for supporting me with a PhD studentship, and Professor Kai Schulze and Professor Michèle Knodt for offering professional opportunities to continue my research endeavours at the Technische Universität Darmstadt while finishing this thesis.

I am furthermore grateful to the members of the PPE (Politics and Policy of the Environment), as well as the EKU (Energie, Klima, Umwelt) reading groups at the University of East Anglia, and the Technische Universität Darmstadt, respectively. Heike Böhler, Dr Viviane Gravey, Brendan Moore, Christoph Priebe, and Dr Tim Rayner generously proof-read and commented on various chapters of this thesis; Heike also offered her expertise to help me generate the Sankey figures.

My two examiners Dr John Turnpenny (University of East Anglia) and Professor Per Mickwitz (Finnish Environment Institute) deserve credit for very carefully reading this thesis, conducting an enjoyable viva and for providing constructive feedback, which has yet again significantly improved the final version of this thesis.

There have been many friends outside my working environment who have helped me in this process—Thomas Barnes has been absolutely wonderful and he also stands for the many I cannot list here. Of course, I would not have been able to do this without strong and ongoing support from my family and especially my parents and my partner Nicole, who have been incredible loving, patient, and encouraging throughout.

To all of you: thank you for all you have done, it means the world to me. You are all important pieces of the jigsaw that makes my PhD.

Abbreviations

ANOVA	Analysis of Variance
CBA	Cost-Benefit Analysis
CPR	Common Pool Resources
CSO	Civil Society Organization
EC	European Commission
EEA	European Environment Agency
EP	European Parliament
EU	European Union
EU ETS	European Union Emissions Trading System
FOCJ	Functional, Overlapping and Competing Jurisdictions
FOI	Freedom of Information
GHG	Greenhouse Gas
IAD	Institutional Analysis and Development Framework
IEEP	Institute of European Environmental Policy
IPCC	Intergovernmental Panel on Climate Change
M	Mean (=average)
MLG	Multilevel Governance
NAAS	Network of Adjacent Action Situations
NAO	National Audit Office
NGO	Non-governmental organization
NPM	New Public Management
OMC	Open Method of Coordination
SD	Standard deviation

UK	United Kingdom
UN	United Nations
UNFCCC	United Nations Framework Convention on Climate Change
USA	United States of America

Figures

Tables

Chapter 1 Introduction

1.1 Climate change and polycentric governance

In an era of rising global greenhouse gas (GHG) emissions (see Jackson et al., 2017) and palpably insufficient policy responses, there is growing concern about whether existing governance systems are capable of dealing with the immense challenge of climate change. The widely perceived failure of the Kyoto-based approach to address climate change by top-down negotiations has precipitated many newer, more bottom-up approaches to climate governance (see Jordan et al., 2015), typified by the adoption of the Paris Agreement in late 2015 (Oberthür, 2016). By that time, scholars had already spent a decade or so exploring governance alternatives to the Kyoto approach (e.g., Lilliestam et al., 2012; E. Ostrom, 2010b; Stewart, Oppenheimer, & Rudyk, 2013; Victor, House, & Joy, 2005). These include for example private and transnational initiatives (Abbott, 2011; Bulkeley et al., 2014), as well as regional or local public policy responses. Taken together, and especially in the area of climate change governance, these activities have increasingly been described as polycentric; that is, spread across many governance levels, actors, and scales (Dorsch & Flachsland, 2017; Jordan et al., 2015; E. Ostrom, 2014b).

According to Elinor Ostrom (2010b, p. 552),

> polycentric systems are characterized by multiple governing authorities at differing scales rather than a monocentric unit […]. Each unit within a polycentric system exercises considerable independence to make norms and rules within a specific domain […].

Note that polycentric governance assumes no central coordinating authority, and yet, her definition still speaks of a 'system.' This way of thinking about governance first entered academic discussions in the 1960s, when US-based scholars started debating the

advantages and disadvantages of different forms of organizing public goods provision, such as policing (V. Ostrom, Tiebout, & Warren, 1961). They searched for ways to coordinate governance activities in the absence of top-down hierarchies or markets (E. Ostrom, 2010a).

But why do some scholars favour polycentric governance, especially in the case of climate change? One argument is that, in contrast to more centralized (or monocentric) governance, polycentric systems have inherently greater flexibility to muster governance responses commensurate with the nature of the problems they seek to address and that they are more resilient to failures in individual parts of the system (E. Ostrom, 2010b). Drawing on such notions, researchers argue that

> The utility of polycentricity in the context of the climate regime is premised upon a theoretical conviction that collective action is more likely than otherwise to take place in a form of small scale networks as they better facilitate face-to-face interactions promoting trust and reciprocity among the actors involved […]. (Lee, Su Jung, & Lee, 2014, p. 33)

Another argument is that polycentric arrangements allow for greater experimentation, which may generate new solutions to unresolved issues—a supposedly key attribute for climate change, whose mitigation still remains an immense challenge (Jordan et al., 2018).

These arguments on the merits on polycentric governance proved especially attractive to climate governance scholars because the Kyoto-based approach had stalled and many emerging efforts to address climate change had in fact already taken a polycentric form (Dorsch & Flachsland, 2017; Jordan et al., 2015; Abbott, 2011; Cole, 2011; 2015; E. Ostrom, 2010c; 2014b; Victor et al., 2005), but were not necessarily designed as such. But even though discussions on polycentric governance have been gaining traction in scholarly and policy-making communities, many of the core assumptions and building blocks of polycentric governance have not yet been fully explored—neither theoretically nor empirically. **This thesis therefore asks precisely**

which factors enable polycentric governance systems to function? In particular, this thesis focuses on evaluation as one such potential factor, whose growing presence has been acknowledged in climate governance debates (e.g. Hildén, Jordan, & Rayner, 2014; Huitema et al., 2011), but which has not yet been sufficiently explored from a polycentric perspective.

The remainder of this chapter proceeds as follows: the next section provides a more detailed overview of polycentric governance. The third section turns to policy evaluation in polycentric governance systems and how to analyse its important but under-specified role. The chapter's last section includes an exposition of the research aims and objectives, and discusses the empirical focus of this thesis, namely climate policy-making in the European Union (EU), which boasts considerably polycentricity and is also an active evaluator of climate policy, and closes with an overview of this thesis.

1.2 Polycentric governance: an overview

Unpacking the role of evaluation in governance from a polycentric perspective first requires clarifying the meaning of the latter. To do so, an analytical distinction between 'polycentricity' and 'polycentrism' is a subtle, but potentially helpful way to understand its inner workings.[2] Linguistically, the term 'polycentricity' is a nominalization of the adjective 'polycentric,' thus according to the Oxford English

[2] The existing literature remains unclear on the use of these different terms and many authors simply use them interchangeably without justification. For example, Aligica (2014) peppers his second chapter with 'polycentricity', 'polycentrism' and 'polycentricism' (see p. 47, third paragraph). Such loose use of terminology risks adding to the "Tower of Babble" that Elinor Ostrom (2006, p. 4) admonished. A review of the three terms with Google Books NGram viewer shows that 'polycentricity' and 'polycentrism' enjoy much wider use than 'polycentricism' (see Appendix 1).

Dictionary connoting "[t]he fact or quality of being polycentric." In other words, 'polycentricity' may be best understood as a descriptor to indicate the apparent structure of governance activities. By contrast, according to the Oxford English Dictionary, 'polycentrism' describes "a situation involving several important elements or powerful parties; [but also] a *system or theory* having or proposing many centres or focal points." But what type of approach is polycentric governance, or polycentrism?

Early on, scholars described polycentrism as a 'concept' (V. Ostrom, 1999a) for theorizing the efficient functioning of metropolitan governments. However, since then, polycentrism has been used in a variety of contexts and for a variety of analytical purposes. As Aligica and Sabetti (2014a, p. 9) write, "[p]olycentricity[3] is a complex multifaceted concept and it is yet to be fully and systematically elaborated as an analytical instrument." The same scholars go on to assert that polycentrism has descriptive, heuristic, explanatory, and normative functions (Aligica & Sabetti, 2014a; see also Jordan et al., 2018). What thus becomes clear is that polycentrism is more than a 'descriptive picture' or metaphor for governance processes, a criticism that has often been levelled at multi-level governance (e.g., Jordan, 2001). Some have even gone as far as identifying and drawing on 'polycentric governance *theory'* (Abbott, 2011), or arguing that the Ostrom enterprise is about "[…] advancing a theory of polycentricity" (Aligica, 2014, p. 38).

Elinor Ostrom herself helpfully distinguishes between frameworks, theories, and models (E. Ostrom, 2007). In a nutshell, frameworks are most general as they specify which elements or variables are relevant in relation to an overall phenomenon (e.g., institutions in Ostrom's work). By contrast, "[…] theories focus on a framework and make specific assumptions that are necessary for an analyst to diagnose a problem, explain its processes and predict outcomes" (E. Ostrom, 2007, p. 25). Last, models "make precise assumptions about a limited set of parameters and variables […]" (E.

[3] Scholars tend to use the terms 'polycentricity' and 'polycentrism' interchangeably. I will use the term polycentrism going forward. However, when citing other scholars, the imprecision in the use of these terms will still appear.

Ostrom, 2007, p. 26). Polycentrism is too narrow to qualify as a general framework for analysis, but it is certainly broader than a model. It thus appears to be a mid-range theory, which seeks to describe and explain some but certainly not all parts of governance processes, so it may best be understood as an emerging governance theory. Over time, corresponding efforts have been underway to refine the initial definition of polycentric governance. One recent example includes McGinnis (2016, p. 5) who writes that

> A **polycentric system of governance** consists of (1) multiple centers [sic] of decision-making authority with overlapping jurisdictions (2) which interact through a process of mutual adjustment during which they frequently establish new formal collaborations or informal commitments, and (3) their interactions generate a regularized pattern of overarching social order which captures efficiencies of scale at all levels of aggregation, including providing a secure foundation for democratic self-governance. (emphasis in original)

Here, we can see how numerous elements have been added to Ostrom's definition, including the structure of the governance system, processes within it, as well as broader normative considerations on democratic governance. Chapter 2 returns to the theoretical building blocks of polycentric governance in greater detail.

Since the 1960s, scholars have been engaging in considerable empirical efforts to test various aspects of polycentrism. Substantial work has focused on common-pool resource (CPR) governance, centring on activities within individual governance centres rather than on the interactions between them (E. Ostrom, 1990). This work has often been conducted on relatively small CPR management systems, such as inland fisheries or forestry (E. Ostrom, 1990; E. Ostrom, 2005). Some scholars have tested their theoretical expectations with quite broad empirical examples, such as the scientific community or the common law system (Tarko, 2017). In more recent work on climate change, Elinor Ostrom drew heavily on the normative and descriptive aspects of polycentrism without clearly specifying underlying causal relationships that may add to or detract from functioning polycentric systems (E. Ostrom, 2010c; 2014b). The

polycentric idea has spurred considerable response from scholars to consider activities in individual centres of governance, but much theoretical and empirical work remains to fully explore the factors that may enable polycentric governance systems to function. This is especially true when it comes to applying the insights from CPR systems to higher levels of governance (see also Singleton, 2017).

Vincent Ostrom long argued that the polycentric idea proves relevant well beyond relatively small scales of governance. He discusses such aspects with a view to federalism in the United States, but also when writing about metropolitan governance (V. Ostrom, 1999a). To date, there remains however a dearth of further theoretical and empirical work to understand exactly how to apply the insights from more local explorations of polycentrism to larger scales of governance. Scholars who study governance at different scales argue that, from an analytic perspective, this kind of up-scaling, also a key element of federal theory, most likely succeeds if problem and solution characteristics are similar across scales regarding the nature of the problem/resource, and the wider social and cultural environment (Gupta, 2008). The extent to which problems and solutions can be scaled (i.e. implemented at another level of governance) has been termed 'transferability,' and when this is not possible, 'transformation' may be needed in order to adjust to particular contexts (Gupta, 2008). In response to such arguments, McGinnis and Ostrom (2008, p. 195) hold that experiences from lower governance levels may indeed be relevant for global governance questions for the following reasons:

1. "The analytical structure of some global problems shares similar features with the analytical structure of many local CPRs [common pool resources].

2. Concepts and tools devised for the analysis of local CPRs provide a solid foundation for building theories and models appropriate for application at a global level.

3. Many global problems (e.g., deforestation) are themselves the result of inadequate solutions at a micro level of a complementary and interactive commons problem."

In a similar vein, Keohane and Ostrom (1994, p. 2) argue that

> Not surprisingly, many of the 'design principles' underlying successful self-organized solutions to CPR [common pool resource] problems appear relevant to the design of institutions to resolve problems of international cooperation as well as those at a strictly local level. For example, both students of local CPRs and of international regimes have identified effective monitoring arrangements as crucial for promoting widespread compliance with rules […]

These arguments suggest that there are theoretical reasons why 'up-scaling' may work, especially with a view to climate change.

What makes polycentrism interesting and relevant to contemporary (climate) governance is the way in which it conceptualizes relationships but also independence among multiple governance centres and the actors contained therein. Importantly, in contrast to hierarchies and networks, different governance centres can move at different speeds without obstructing overall progress. Elinor Ostrom thus suggests that this may help to avoid gridlock, which she claimed had long plagued the predominantly top-down international approach to climate governance (E. Ostrom, 2010c; 2014b). At heart, the idea of a polycentric governance 'system' revolves around interacting but independent governance centres, which sometimes compete and/or learn from each other, and that this competition and learning can under certain circumstances produce substantially 'better' policy outcomes than hierarchical systems (V. Ostrom et al., 1961). As far as individual governance centres in polycentric systems are concerned, Elinor Ostrom (2005, p. 259) proposes a number of revised design principles to ensure effective organization for common pool resource governance within them, namely "clearly defined boundaries, proportional equivalence between benefits and costs, collective choice arrangements, monitoring, graduated sanctions, conflict resolution mechanisms, minimal rights to organize [and] nestled enterprises." While she acknowledges that even if these conditions are met results may not always be positive because some actors simply do not self-organize or because small governance centres have only limited scientific possibilities (E. Ostrom, 2005, p. 282), even fewer theoretical insights are

available regarding how governance centres interact productively to form a polycentric 'governance' system (and whether indeed we can speak of a system at all). This thesis focuses on how evaluation may foster polycentric governance in the case of climate change.

1.3 The role of evaluation in polycentric governance systems

A system of polycentric governance would amount to more than the sum of its parts by capitalizing on synergies that arise from its systemic aspects, such as mutual learning among the governance centres within the system. In order to effect the shift from 'polycentricity' to 'polycentrism,' or from a polycentric form to a polycentric *governance system*, scholars have advanced the normative claim that formally independent governance centres *have* to interact to a certain degree. In their widely cited article, Vincent Ostrom, Tiebout and Warren (1961) recognized this by highlighting that

> To the extent that they [governance centres] take each other into account in competitive relationships, enter into various contractual and cooperative undertakings or have resource to central mechanisms to resolve conflicts, the various political jurisdictions in a metropolitan area may function in a coherent manner with consistent and predictable patterns of interacting behavior [sic]. To the extent that this is so, they may be said to function as a 'system.' (p. 831)

Simply paying attention to activities and experiences in other governance centres will not necessarily generate a polycentric system because after all, governance actors must act on this information in one way or another, otherwise it would be futile to speak of a 'system.'

Only if governance centres interact can we properly speak of a polycentric 'system' or the emergence of polycentrism (see above). So how do they interact and what role may evaluation play in this process? In her writing, Elinor Ostrom implicitly assumes that knowledge flows between governance centres. For example, she writes that

"[a]s more information is provided about these small-scale, but cumulatively additive, benefits [of emissions reducing activities], one can expect further efforts to be undertaken that cumulatively and significantly reduce GHG emissions" (E. Ostrom, 2010c, p. 553). But she falls short of exactly specifying from *where* and from *whom* this information is thought to emerge, aside from general references to higher governance levels (E. Ostrom, 2010c), and she does not explicitly address and discuss evaluation. Other scholars have gone somewhat further. For example, Abbott (2011, p. 586) writes with a view to enabling interactions that

> Information and networking schemes […] are particularly important in this regard […] *if cities, firms, CSOs [civil society organisations], and other actors are to observe their peers on a global scale, benchmark their strengths and weaknesses, and learn from their successes and failures, schemes that facilitate interaction, disseminate information, and encourage learning are essential.* (emphasis added)

But again, precisely *from where* and *how* the information emerges remains unclear. In the world of public (i.e. state) policy, can *ex-post* (i.e. retrospective) evaluation be a vehicle to generate systematic and evaluative insights on the functionality and effects of policy interventions in particular governance centres in a polycentric system? While policy-makers may of course use other ways to learn from and/or 'evaluate' policies, such as their own personal perceptions or ideological preferences, evaluation can in principle provide systematic and in-depth analysis of the effects of individual (climate) policies. However, evaluation may also be subject to the same political pressures that apply in other aspects of policy-making, hence its existence and especially its quality should not be assumed (Bovens, Hart, & Kuipers, 2006).

Are there consequently certain characteristics of policy evaluation that may be particularly paramount in order to facilitate polycentric governance? Polycentric governance theorists have long argued that monitoring and enforcement often works better in the hands of localities rather than in a centralized fashion (E. Ostrom, 2010b; 2014b; see also Cole, 2011), but they fall short on discussing the specific characteristics of evaluations in polycentric systems. Symptomatically, there has to date only been

minimal engagement between emerging literatures on polycentric governance and policy evaluation.

A first crucial step in specifying evaluation's potential role in polycentric governance (see Chapter 2) is to complement the discussion of the latter with (an operational) definition of evaluation. As Furubo, Rist and Sandahl (2002, p. 2) admit, "[i]f asked for a definition of evaluation, the attempted answer might be seen as a never-ending story" (see also Pawson & Tilley, 1997). Evaluation has generally been described as "[…] the process of determining the merit or worth or value of something; or the product of that process" (Scriven, 1981, p. 53). Thus, it is undeniably a normative endeavour (Fournier, 2005; Vedung, 1997). But not all definitions recognize this. Consider, for example, the OECD (2002), which defines evaluation as "[t]he systematic and *objective* assessment of an on-going or completed project, programme or policy, its design, implementation and results" (p. 21; emphasis added). Such claims to objectivity, however, contradict those who argue that evaluation is fundamentally value-based. Others thus define evaluation more broadly. For instance, the *Encyclopedia of Evaluation* sees it as

> […] an applied inquiry process for collecting and synthesizing evidence that culminates in conclusions about the state of affairs, value, merit, worth, significance, or quality of a program, product, person, policy, proposal, or plan. Conclusions made in evaluations encompass both an empirical aspect (that something is the case) and a normative aspect (judgment about the value of something). It is the value feature that distinguishes evaluation from other types of inquiry, such as basic science research, clinical epidemiology, investigative journalism, or public polling. (Fournier, 2005, p. 139-140)

This definition is relatively broad and encompasses many different potential targets of evaluation; furthermore, it makes a clear distinction between basic science research and evaluation.

But what about (public) *policy* evaluation? Crabbé and Leroy (2008) explain that a definition of *policy* evaluation hinges in part on what we mean by 'policy' or the 'policy

process.' They propose three fundamental views of policy, which envision different roles for policy evaluation: those who envision policy as a rational process of problem-solving see evaluation as a 'control loop' or correcting device that indicates how well a policy 'solves' the problem it seeks to address. However, as the idea of entirely rational policy-making has proven increasingly problematic, policy has also been conceptualized as a 'political' or 'discursive' interaction (see also Fischer, 2006). Policy then becomes the result of interactions between actors who exercise their power in order to further their interests – and evaluation thus a way of understanding these interactions (Crabbé & Leroy, 2008). Finally, a policy may be seen as an 'institutional phenomenon' where institutions structure and stabilize human behaviour; thus the task of evaluation broadens even further to include a study of these institutions (Crabbé & Leroy, 2008). Underlying this argumentation is that policy, and in particular *public* policy is a course of action that is typically undertaken by public or governmental actors – though of course, there are numerous broader meanings of policy that will not be considered here (Hill & Varone, 2017). While the concept of policy remains fuzzy, this thesis conceptualizes it as a course of action by governmental actors in order to solve some perceived problem.

Numerous authors emphasize that evaluation is also a purposeful process, geared towards providing insights for policy makers. Accordingly, the prominent scholar Evert Vedung (1997, p. 3) defines *policy* evaluation as a

> careful retrospective assessment of the merit, worth, and value of administration, output and outcome of government interventions, which is intended to play a role in future practical action situations.

Along very similar lines, Crabbé and Leroy (2008, p. 1) write that policy evaluation

> is a scientific analysis of a certain policy area, the policies of which are assessed for certain criteria, and on the basis of which recommendations are formulated.

The latter definition is useful, because it clearly delineates evaluation as a 'scientific analysis' – in other words not a position or advocacy paper by an interest group or a newspaper article – and it also explains that evaluation comes with a purpose. But not all definitions of evaluation include a purpose. For example, Fischer (2006) writes that "[p]olicy evaluation is […] the activity of applied social science typically referred to as 'policy analysis' or 'policy science' (p. 2). Clearly, this definition focuses on the process of evaluation, rather than its outcome.

This thesis broadly follows Vedung's (1997) definition (see above) given the public policy focus of this thesis, but empirically it casts a somewhat wider net as suggested by Huitema et al. (2011), who argue that producing recommendations for policy-makers is certainly an important, but not a necessary characteristic of policy evaluation. Policy evaluation may, as Crabbe and Leroy (2008) point out, be conducted for other purposes and by non-state actors for many different target audiences, some going well beyond policy-makers (Huitema et al., 2011). The efficacy of evaluation-based recommendations is further questionable given that the uptake of (evaluation) knowledge or ideas in policy is often not straightforward and may take long to manifest (Johnson et al., 2009; Radaelli, 1995). Finally, in line with Scriven's (1981) definition, this thesis understands and studies evaluation both as a 'process' and as an 'outcome.' Evaluation is a process because it is in many ways a practice, which often involves the interaction of numerous actors who together decide what is of value and what is not (see the second definition of policy above). However, in line with Huitema et al.'s (2011) approach, frequently the outcome of this process are evaluation reports (hereafter 'evaluations'), which in turn become the backbone of the empirical analysis in this thesis.

1.4 Analysing policy evaluation

Policy evaluation has been debated in academia and elsewhere ever since it rose to prominence more than five decades ago (Stame, 2003). In these debates, the EU and its constituent parts have served as important loci of evaluation activity (Stern, 2009;

Summa & Toulemonde, 2002; Toulemonde, 2000). However, most of the corresponding evaluation literatures consist of prescriptive 'how to' guides for evaluations and they tend to be practitioner rather than theory-led. In other words, by extension of Hill's (2017) argument on policy analysis, much of the existing literature focuses on 'evaluation *for* policy' rather than analysis *of* evaluation. In a 'field' that is already heavily infused with normative elements, this development may risk that evaluation theorists speed away from the realities that evaluators encounter in their daily work and the evaluation they practice. There is some evidence that this may be happening – for example, Christie (2003) found that evaluators were at best partially informed by normative and prescriptive evaluation theory. Thus, no less than the former President of the American Evaluation Association, Debra Rog, argues that "[w]e need more study of evaluation practice itself; we need to accumulate knowledge about evaluation" (Rog, 2012, p. 38). This is especially the case for the role of evaluation in polycentric governance.

In order to understand how policy evaluation in the EU may facilitate polycentric governance it is thus wise to take 'analysis *of* policy evaluation' as a starting point. As Hill (2017, p. 5) argues, in principle, "[e]valuation marks the borderline between analysis *of* policy and analysis *for* policy" because it "[…] may be either descriptive or prescriptive." A smaller, but emerging literature focuses on analysis *of* evaluation (Segerholm, 2003; see also Hogwood & Gunn, 1984, p. 228), and there have been repeated calls to do more of this, including in the context of social science and political science theory (Duscha, Klemisch, & Meyer, 2009; Vo & Christie, 2015). King (2003, p. 57-60) identifies six reasons why much less work has been done to test evaluation theory so far:

- Lack of conceptual consensus
- Practical focus
- Continuing focus on evaluation models and methods
- Focus on program theory
- Lack of research support
- Relatively young field

According to King (2003), the 'lack of conceptual consensus' originates from the interdisciplinary nature of the evaluation field, whereas the practical focus of many evaluators has afforded little room for them to critically analyse their own theory and practice. Finally, there tends to be little funding available for studying evaluation, in part because this is still a relatively new field. These reasons notwithstanding, there is some research on the politics of policy evaluation, as well as the use of evaluation theory and methods in different settings (see Chapter 2). For example, Segerholm (2003) urges researchers to study evaluation within its wider political and organizational context and suggests using the 'evaluation cycle,' which comprises evaluation initiation, implementation, results, utilization, as well as the evaluation context and evaluation's theoretical orientation as a framework in such studies. This is particularly important given that governance arrangements tend to shape policy evaluation by for example demanding that evaluation fulfil different roles and policy evaluation may in turn become a critical element in governance debates and arrangements (Gore & Wells, 2009; see also Schoenefeld & Jordan, 2017).

In a similar vein, Radaelli and Dente (1996) argue that the role of policy evaluation can only be understood after analysing how policies are made—and show how early evaluation theory ('the age of innocence') relied on linear and rational conceptualizations of policy-making, and thus an instrumental and direct role for evaluation knowledge, while later theories began to engage with the incremental, evolutionary, and often highly political nature of policy processes, and thus foresee a different role for evaluation, such as providing new ideas that percolate slowly into policy-making, or a moderator role for evaluators. Attention to 'systemic' factors also emerged in the evaluation literature, for example in an edited collection by Rist and Stame (2011), who argue that individual evaluation studies are increasingly insufficient and that it is therefore necessary to understand evaluation from a broader and systemic perspective. In sum, there is a drive to understand evaluation activities more broadly.

This thesis endeavours to contribute to such analysis *of* evaluation in polycentric settings. Lack of analysis *of* evaluation goes hand-in-hand with similarly low levels of interaction between the evaluation and (polycentric) governance literatures (Segerholm,

2003). For instance, Hanberger (2012, p. 10) writes that "[t]ere is clearly a dearth of knowledge regarding how M&E [monitoring & evaluation] work and function in different models of governance." This also includes evaluation activities in the context of the nation state (Segerholm, 2003). But will more knowledge in this domain generate benefits because "[i]f more attention is paid to the governance structure in which evaluation is embedded, we can arrive at a better understanding of the implications of evaluation in public policy and governance […] (Hanberger, 2012, p. 10)?

Some of this work has already begun because changing governance arrangements in multiple contexts have not entirely escaped the attention of evaluation theorists. For example, Hertting and Vedung (2012) explain how some elements of existing evaluation approaches may be useful to evaluate policies enacted through governance networks, while some new elements may be needed to cater to existing governance debates (e.g., accountability). However, as Gore and Wells (2009, p. 161) argue,

> […] while the interest in governance by the evaluation community is to be welcomed, it has to date confined itself either to issues of evaluation method (e.g., participatory methods), to using evaluation to contribute better to policy making (e.g., democratic evaluation), or to designing evaluation frameworks (e.g., to take account of new governance arrangements). There appears a continued absence of more theoretically informed work which sets out, for example, how issues of power, resource dependency, ideas, and networks shape policy outcomes.

Whilst some scholars have studied policy evaluation across the EU's multilevel environment (e.g., Mickwitz, 2013), surprisingly little has been written about evaluation's role in increasingly polycentric governance settings. In fact, even as recently as 2015, Jacob et al. argue that there is "[…] little systematic comparative research across countries… and this body of research is still at a relatively early stage" (p. 2). On the one hand, one may argue that multiple centres of governance need standardized evaluation practices and metrics in order to compare their policy outcomes (see Duscha et al., 2009; V. Ostrom, 1999b). However, efforts to standardize policy evaluation systems may go against the very essence of polycentric governance (see E.

Ostrom, 2014b). Taken together, it thus remains an open question whether highly uniform evaluation practices can deliver on such diverging policy needs or whether a more decentralized form of evaluation would be more useful (Schoenefeld & Jordan, 2017).

In recent years, scholars have developed important analytical categories in order to analyse policy evaluation. For this thesis, an absolutely central way to look at policy evaluation from a polycentric perspective is to distinguish between *formal* (i.e. state-driven) and *informal* (i.e. society-driven) evaluation (Hildén et al., 2014; Schoenefeld & Jordan, 2017). Polycentric governance scholars have time and again emphasized that states are not the sole sources of governance – or evaluation - (see Chapter 2), and thus counsel to look at *both* state and non-state actors, especially in order to assess the capacity of different governance actors to self-organize. Evaluation theorists, too, have for several decades written on the nature of different evaluation actors, and especially considered them with a view to their independence from the state (see Chapter 2). If we can establish that both formal and informal evaluations constitute important factors in facilitating polycentric governance, then the relationship between them also matters. Are the contributions of formal *and* informal evaluations to polycentric governance unique and complementary or are they similar and perhaps overlapping? Looking at evaluation from the polycentric perspective in effect means to identify how far evaluation facilitates polycentric climate governance in the EU, but also to be open to the possibility that evaluation *itself* may have polycentric characteristics (Schoenefeld & Jordan, 2017). This is because in line with the arguments on *state* and *non-state* actors above, evaluation is certainly not limited to different public actors, but also includes private or third-sector actors. Nevertheless, we have so far only a very limited understanding of what their respective contributions via evaluation may be. Therefore, the distinction between formal and informal evaluation is a core way to study evaluation in this thesis.

1.5 Overview of this thesis

The central aim of this thesis is to understand precisely which factors enable polycentric governance systems to function, with a particular focus on the potential and actual role of policy evaluation. This aim translates into two specific objectives:

Objective 1: Identify the key foundational ideas of polycentric governance theory and relate these to relevant debates on policy evaluation in order to understand the *potential* role of evaluation in facilitating climate governance.

Objective 2: Test these theoretical expectations in the case of the European Union in order to understand the *actual* role of evaluation in climate governance.

Empirically, this thesis focuses on climate change policy-making in the European Union, which exhibits considerable polycentricity in its approach to governance. The EU does not have a single locus of authority and decision-making, but rather multiple routes through which different actors, most importantly Member States and the EU's main institutions (Commission, Council of Ministers, European Parliament and the European Court of Justice) decide on and conduct (climate) governance (Peterson & Shackleton, 2012; T. Rayner & Jordan, 2013). Scholars have pointed to "the EU's inherent polycentricity—i.e. its active encouragement of experimental efforts at multiple levels, with active steering of actors at local, regional, and national levels […]" in climate change policy-making (T. Rayner & Jordan, 2013, p. 75) as a source of its strength (see also M. D. McGinnis, 2016; E. Ostrom, 2010c). This structure did not emerge completely by accident – in fact, the "EU's 'founding fathers' deliberately set out to prevent power from accumulating in ways that had dragged Europe into two world wars" (Jordan, Van Asselt, Berkhout, Huitema, & Rayner, 2012, p.46). But even though European institutions were certainly consciously designed and shaped, they have also taken on a life of their own, especially because the Commission has always had the exclusive right to propose policies (see Peterson & Shackleton, 2012). Climate change policy has often been an area that is understood to work particularly well – a system where international policy leadership appears to emerge from a 'leaderless,' polycentric system (T. Rayner & Jordan, 2013).

For example, while the EU as a whole has not been able to agree on firm long-term emissions reduction targets beyond 2030—it only has political targets until 2050 (see Dupont & Oberthür, 2015), the UK's Climate Change Act contains a much longer-term legally binding emissions target by 2050 (Benson & Lorenzoni, 2014), and other states have other preferences (Jordan, Huitema, Van Asselt, Rayner, & Berkhout, 2010; Oberthür, 2016). Numerous actors in the EU including its Member States, institutions at the EU level and sub-national actors have been deploying policies in order to address climate change for well over thirty years (Jordan et al., 2010)—so much so that the number of individual climate policy instruments from Member States alone had swollen to over 1,300 by 2013 (Schoenefeld et al., 2018). But what counts as a (public) climate policy? The answer to this question is by no means a trivial matter, because, as Feldman and Wilt (1996, p. 63) explain, "[…] there is a practically limitless range of activities that may be counted as having an impact on global climate change." The difficulty is compounded by the fact that what constitutes a policy is also contested, as public policies are often related to a range of (non)decisions by governmental actors that may change over time (see Hill & Varone, 2017). Drawing on Hill (2017), this thesis therefore defines climate policies as courses of action undertaken by governmental actors in order to reduce greenhouse gas emissions.

This thesis builds on the notion that governance arrangements influence policy evaluation and that policy evaluation has the potential to become an important element of governance (Gore & Wells, 2009). Climate change policy in the EU is a suitable setting to explore such dynamics, because it exhibits much 'polycentricity' and because the EU is also an active evaluator of environment (European Environment Agency, 2016; Mickwitz, 2013) and especially climate policy (Haug et al., 2010; Huitema et al., 2011; Schoenefeld & Jordan, 2017). This thesis focuses on three key governance centres in the EU, namely the EU governance level, as well as Germany and the UK (both national level only). The latter two are not only the two top-emitters of greenhouse gases in the EU (and thus highly relevant climate change policy actors), but importantly for this thesis they are also known for their efforts to address climate change and to evaluate their policies (Derlien, 2002; Gray & Jenkins, 2002; Jacob, Speer, & Furubo, 2015; Wurzel & Connelly, 2011). As the two largest governance centres in polycentric climate

governance in the EU, they are thus highly relevant places to explore the role of policy evaluation.

The remainder of this thesis proceeds as follows. Chapter 2 provides a detailed review of polycentric governance theory with a view to a theoretical role of policy evaluation in facilitating climate governance. It distils three foundational ideas of polycentric governance and relates these ideas to relevant debates in policy evaluation literatures. Note that the extensive debates of (evaluation) knowledge utilization (e.g., Johnson et al., 2009; Rich, 1991) remain, by and large, outside the scope of this thesis. In addition, while the focus of this thesis is *ex-post* (i.e. retrospective) climate policy evaluations, literatures on environment and climate policy evaluation in the EU context suggest that *ex-ante* (prospective evaluation of future policy impacts; often termed 'impact assessment' (Radaelli, 2010; Turnpenny, Russel, Jordan, Bond, & Sheate, 2016)), *ex-nunc* (monitoring ongoing policy) and *ex-post* evaluation (Crabbé & Leroy, 2008) may at times be used together or not be clearly distinguishable in theory and practice. Relatedly, Chapter 3 explains how the data that EU Member States collect on climate policies includes aspects of *ex-ante*, *ex-nunc*, and *ex-post* data (Hildén et al., 2014; Schoenefeld et al., 2018). In doing so, Chapter 3 draws on the theoretical review in order to uncover important research gaps on climate policy evaluation at the EU level, as well as in Germany and in the UK, which are some of the most productive sites of climate policy and corresponding evaluation.

Chapter 4 describes the research methods employed in this thesis, including the novel coding scheme for analysing evaluations and the new database of climate policy evaluations from the aforementioned jurisdictions. Importantly, this thesis focuses on climate mitigation (i.e. it does not consider adaptation policy). But the scope nevertheless includes a broad range of policies with a view to those that states report as climate policy to the United Nations Framework Convention on Climate Change (UNFCCC)—(see Huitema et al., 2011; Schoenefeld et al., 2018). Chapters 5 and 6 present the coding results, namely on state-based (*formal*) and society-based (*informal*) climate policy evaluations respectively and their key characteristics with a view to facilitating polycentric governance. Chapter 7 presents an empirical comparison of

formal and informal evaluations. Chapter 8 contains a detailed, theoretical analysis of the empirical findings. Finally, Chapter 9 concludes the thesis with a reflection on how evaluation could in theory and how it actually facilitates polycentric climate governance in the EU, policy recommendations, and some ideas for future research in this important topic area.

Chapter 2 Policy Evaluation in Polycentric Governance Systems

2.1 Introduction

A key starting point for polycentric governance scholars is the idea of heterogeneity in governance, which the Oxford English Dictionary loosely defines as "composed of diverse elements or constituents." "The Ostroms pointed toward heterogeneity, diversity, context, and situational logic as critical elements in the analysis of institutions, governance, and collective action" (Aligica, 2014, p. 5). While heterogeneity can take many forms, referring to diversity in capabilities, preferences, beliefs, information, but also social, cultural or linguistic aspects (Aligica, 2014, p. 4-5), this chapter focuses on the conceptual consequences of such heterogeneity for theorizing the role of policy evaluation in the shift from polycentricity to polycentrism (see Chapter 1).

To do so, this chapter provides an overview of polycentric governance theory in terms of positivism, normative elements and key variables. It then disentangles three foundational ideas or assumptions on which polycentrism builds, namely that context matters in governance, that actors can and do self-organize in order to address pressing governance challenges and that governance centres, while independent, interact to fully generate the hypothesized benefits of polycentric governance. The chapter explains the origins of these ideas, specifies their theoretical implications for polycentric governance and draws together key existing empirical research. This is done in order to assess how the idea of monitoring currently features in work on polycentrism, and how related key insights may be developed in order to analyse what role evaluation may play to contribute to the shift from polycentricity to polycentrism with a view to the foundational insights. To do so, the chapter ultimately combines polycentric governance

and policy evaluation literatures in novel ways in order to advance polycentric governance and policy evaluation theory. The chapter concludes by further elaborating and specifying the key empirical research gaps, which Chapter 1 already flagged, and which will be addressed in the chapters that follow.

2.2 Positivism, key variables, and normative theory

From the outset, and as Chapter 1 recognised, it is critical to appreciate that theory on polycentrism contains a subtle blend of both normative and positive elements (Aligica, 2014; M. D. McGinnis, 2016). In his early and later re-published work, Vincent Ostrom explains that polycentrism does not simply provide an explanation of the status quo, but is rather a theory which is capable of making normative prescriptions, such as identifying necessary conditions for polycentrism to work (V. Ostrom, 1999a). In more recent contributions, the normative element has become even more pronounced, as for example McGinnis and Ostrom (2012) specify what "polycentric governance *requires*" (p. 15; emphasis added). In sum, the polycentric approach must be understood as both a positive and a normative project.

Such normative considerations should however not obscure the extraordinary amount of empirical work that scholars in and around the *Ostrom Workshop in Political Theory and Policy Analysis*[4] have been undertaking for at least five decades in order to gauge the (normative) polycentric approach against empirical realities. Elinor Ostrom's book on *Governing the Commons* (1990) draws on vast empirical evidence to validate and further develop key polycentric ideas. Her 'design principles'[5] for common pool

[4] http://www.indiana.edu/~workshop/

[5] Elinor Ostrom's design principles for successful CPR governance (E. Ostrom, 1990, p. 90):

1. Clearly defined boundaries.
2. Congruence between appropriation and provision rules and local conditions.
3. Collective-choice arrangements.
4. Monitoring.

resource governance systems have since found further empirical support around the world (E. Ostrom, 2005). Given decades of empirical work and particularly Elinor Ostrom's affinity for 'grounded research' and interdisciplinarity (E. Ostrom, 2005), it would be inappropriate to relegate polycentrism to the realm of purely normative governance theories. By the same token, polycentrism clearly contains normative elements, which will likely become stronger as polycentrism gains traction and application on numerous issues, including climate governance. In fact, Aligica (2014) argues that Elinor and Vincent Ostrom have moved from empirical explorations towards more normative elements over time. In fact,

> Certain normative assumptions and preferences are undoubtedly and inescapably embedded at a very basic and intuitive level in the perspectives advanced by scholars, like the Ostroms, who explore collective action and institutional arrangements. (Aligica, 2014, p. 17)

This thesis thus endeavours to make these normative elements explicit and engage with them in the context of studying policy evaluation.

The presence of normative aspects in polycentrism derives from the Ostroms' general scholarly approach, which seeks to elevate theory over methods (Aligica & Sabetti, 2014a, p. 2). This approach reacts to the positivist doctrine starting in the 1960s, where scholars endeavoured to build theory starting from empirical insights (E. Ostrom, 2014a). By contrast, Elinor Ostrom (2014a) advocates that "[…] the development of theory precedes the choice of appropriate methods to test a theory" (p. 218). She furthermore elaborates that "[…] theory has also come to mean for many political scientists a set of logically connected statements without the requirement that assumptions used in a theory have themselves *already* been established as empirical laws" (p. 218; emphasis in original). However, reverting back to an earlier point, Elinor

5. Graduated sanctions.
6. Conflict resolution mechanisms.
7. Minimal recognition of rights to organize.
8. Nested enterprises (for common pool resources that are part of larger systems).

Ostrom also had a strong affinity for empirical work. Consequently, she argues that while "theory precedes empirical work […], empirical studies help to refine our theoretical understanding of the world […]." (E. Ostrom, 2014a, p. 222). Taken together, Elinor Ostrom thus advocates a dialectic relationship between theory and data, but an approach that allows normative elements because theory comes before empirics. This general stance may in part explain the presence of normative elements in scholarship on polycentrism.

There has of course been a strong movement in political science and related fields to develop context-independent and generalizable theory. As Benjamin (1982, p. 69) argues,

> During periods of relative social-economic and political stability, social scientists are lured into a false sense of security regarding the ahistorical validity of empirical generalizations.

Thus, if social conditions are ever changing and unstable, Benjamin (1982, p. 93) holds that

> [t]he continual need to develop, question, and reformulate theory (the general structuring principles that allow a temporary but necessary ordering of the political and social processes) should now be considered the most important element of the logic of inquiry on which to concentrate. If one grants this point, then the context, assumptions, conceptualization, and reconceptualization of the way the questions are formulated takes on crucial significance.

According to Austen-Smith and Banks (1998, p. 259), a "[p]ositive political theory is concerned with understanding political phenomena through the use of analytical models which, it is hoped, lend insight into why outcomes look the way they do and not some other way." These models typically include assumptions such as rational individuals or the way individuals interact in game-theoretic situations (Austen-Smith & Banks, 1998). While polycentrism provides a normative panoramic vision of

the governance landscape, many of the inner workings – both in normative and empirical terms – have yet to be fully explored.

2.3 Polycentrism – three 'foundational' ideas

Building on the underlying ideas of heterogeneity in governance (see above), the polycentric governance approach flows from and finds support in three foundational ideas. The first foundational idea of polycentrism is that polycentric governance emerges precisely because actors at various levels have the capacity and, given adequate circumstances, the willingness to self-organize. In earlier writings, Vincent Ostrom has pointed to the "self-organizing tendencies" of such actors in polycentric systems (V. Ostrom, 1999a, p. 59). In order to self-organize, (new) actors need governance systems that are sufficiently open and flexible, a sense that they have some capacity to affect and change the rules to which they are subjected, and a feeling of motivation to actively participate in enforcement (V. Ostrom, 1999a). In this process, self-organizing actors may thus benefit from sufficient, place-sensitive information on previous climate policies that is readily available and accessible (see above). If these conditions are met and actors self-organize, outcomes may be 'better' than top-down solutions.

Polycentric governance theory holds that this is because those who have knowledge of the particular 'local' governance context tend to be better placed and willing to make rules and regulate their own behaviour. In recent decades, empirical evidence from common pool natural resource management literatures has built up to emphasize this point. Crucially, the assumption that actors will always deplete common pool resources in the absence of coercion from a higher authority (Hardin, 1968) does not withstand empirical scrutiny across all cases (E. Ostrom, 1990), although Elinor Ostrom very much recognizes potential drawbacks of polycentric governance arrangements, such as the possibility for free riding and potential under-provision of public goods (E. Ostrom, 2010c). In fact, across multiple natural resource types including fisheries, and water or timber production, local actors managed to build enduring institutional systems to self-govern their local resource use (E. Ostrom, 1990). Thus, in some cases actors appear to exhibit the capacity to self-organize and

outperform top-down solutions. This thesis assesses to what extent this proposition materializes in the case of climate policy evaluation.

The second foundational idea is that context matters and that no rule or policy will produce effects irrespective of their wider context (Aligica, 2014). Elinor Ostrom and others conceptualize the influence of 'context' through the Institutional Analysis and Development (IAD) Framework. According to McGinnis (2011, p. 51),

> The IAD framework contextualizes situations of strategic interaction by locating games within social, physical, and institutional constraints and by recognizing that boundedly rational individuals may also be influenced by normative considerations.

This line of reasoning underpins one of Elinor Ostrom's key messages, namely that there are no policy 'panaceas' that will hold in all situations irrespective of the context (E. Ostrom, Janssen, & Anderies, 2007). Different contexts require different approaches as there is no one-size-fits-all approach.

This insight has long been acknowledged in international climate governance. In 1992, the United Nations Framework Convention on Climate Change (UNFCCC) stated that in order to address climate change, "[…] policies and measures should take into account different socio-economic contexts […]" (Article 3[3]). In consequence, it is only by paying close attention to the context that analysts can understand how actors and rules generate particular effects (Aligica, 2014)—and by extension policy evaluation should, therefore, also be place and time specific. Furthermore, because context and 'local' conditions matter, multiple solutions at various governance scales including many actors may thus generate 'better' outcomes than a single, hierarchical approach. This is one of the most central ideas of polycentrism. However, "[n]o *a priori* judgment can be made about the adequacy of a polycentric system of government as against the single jurisdiction" (V. Ostrom et al., 1961, p. 838). The effectiveness of polycentric governance depends, at least in part, on its fit with the wider context into which it is placed. Building an understanding of the successes and failures of polycentric governance systems thus requires close attention to context—including in (public)

policy evaluation. Therefore, learning across contexts requires intimate knowledge of contextual variables—including historical, geographical, cultural or ideational aspects to name but a few (Aligica & Sabetti, 2014b).

The third foundational idea holds that if polycentrism is to emerge, governance centres need to interact, but without generating strong interdependencies. But what is a 'governance centre'? Scholars in the polycentric tradition differ in their understanding. For example, Elinor Ostrom (2012, p. 355) writes that "[a] polycentric system exists when multiple public and private organizations at multiple scales jointly affect collective benefits and costs", thus taking an 'organization' as the core unit of analysis. In a slightly different way, Vincent Ostrom, Tiebout and Warren (1961, p. 831) write about "centers [sic] of decision-making" as the core unit, with less emphasis on 'organizations.' In a different vein, Elinor Ostrom (2005, p. 257) stresses that "complex, polycentric systems of governance that are created by individuals", thus focusing on people. In other places in the same book, Elinor Ostrom (2005, p. 269) writes about "the presence of governance activities organized in multiple layers of *nested* enterprises" (emphasis added). These differing definitions show that what constitutes a 'governance centre' is by no means clear, as it may range from individuals to all types of organizations or enterprises all the way to more fuzzily described 'centres of decision-making.' To complicate things more, a recurring theme in Elinor Ostrom's scholarship is that governance centres are 'nested' (see quote above), which creates the challenge to not only tell governance centres apart in a horizontal, but also in a vertical, way and to understand their potential linkages. Looking across the relevant literatures, the ideas of 'decision-making' and 'independence' run quite deeply and are probably theoretically more relevant than the exact nature of the organization (or the number of people involved) that make up a governance centre. This thesis defines governance centre in a broad sense, that is, as any organization or organizational unit that has authority to make some decisions and is reasonably independent in doing so (see V. Ostrom et al., 1961). This definition therefore encompasses the level of the nation state and supra-national organizations like the EU.

Linked to the idea of 'nesting,' what drives interactions between centres of governance in polycentric systems? There are numerous potential mechanisms. Vincent Ostrom believed that governance centres will interact more or less automatically if they have sufficient incentives to do so (V. Ostrom et al., 1961). Overlapping jurisdictions may be one reason why centres interact. For example, writing on the IAD, McGinnis (2011, p. 52) proposes that interaction may take place through a "network of adjacent action situations" (NAAS) where individuals or organizations simultaneously participate in multiple rule-making venues in a polycentric system. These individuals or organizations become bridges between different governance centres to foster interaction. In other cases, interaction may emerge because of market-like competition—for example, when different governance centres offer the same service. If two municipal governance entities provide the same service, people are likely to choose the one that they see as most favourable, depending on the dimension that matters most to them (e.g., cost; quality of the service, etc.). However, scholars from other fields have proposed a range of additional mechanisms. For example, policy diffusion and transfer scholars distinguish between learning, competition, coercion and mimicry as forms of interaction (Marsh & Sharman, 2009). While multiple disciplines have identified these kinds of mechanisms, scholars differ significantly on which mechanisms matter more and, importantly, how much external stimulus may be required to stimulate interaction. By definition, the polycentric approach excludes ideas around top-down coercion, as governance centres are a priori thought to be independent.

In climate change governance, the threat of 'carbon leakage' provides one potentially strong (external and market-driven) incentive for governance centres to experiment with reducing their carbon dioxide emissions efficiently and potentially cooperatively. Carbon leakage generally refers to the idea that actors may shift activities that cause carbon pollution from jurisdictions with more regulation to those with less in a classic 'race to the bottom' (E. Ostrom, 2014b). Thus, if public policy-makers perceive carbon leakage as a threat—such as heavy industry moving to other countries, with corresponding job losses—they may have significant incentives to identify the least intrusive ways to reduce carbon emissions and ensure that other governance centres take equivalent action.

An additional reason to look beyond one's own governance centre is to learn from the successes and failures of others, especially because policy-makers tend to be risk-averse (Howlett, 2014). While the concept of policy learning has been subject to much scholarly debate, multiple authors point to learning as some change in behaviour or beliefs, following the impact of experience, new information, or changing circumstances (Bennett & Howlett, 1992). Of particular interest to this thesis is 'lesson drawing,' which is one form of learning that focuses on using the 'lessons' or experiences from one governance context in another (Rose, 1991; 1993). Thus, Rose (1991) explains that

> A lesson is here defined as an action-oriented conclusion about a programme or programmes in operation elsewhere; the setting can be another city, another state, another nation or an organization's own past. (p. 7)

Crucially, rather than being compelled by some top-down authority, "lesson-drawing tends to be voluntaristic" (Rose, 1991, p. 9) and thus fits well with ideas on polycentric governance. Climate policies may, for example, generate politically desirable side-effects, such as improvements in human health or reducing congestion (T. M. Thompson, Rausch, Saari, & Selin, 2014). Learning about experiences with such (beneficial) side effects and their political consequences may thus be another incentive to seek information about experiences in other governance centres. Lesson-drawing is not the 'normal' state of affairs, but rather emerges from an underlying level of 'dissatisfaction' with the status quo that prompts a search for lessons from elsewhere (Rose, 1991). The aforementioned risk aversion among policy-makers may be one such source of 'dissatisfaction.' In the area of climate change, where there are currently no examples of far-reaching policy success in addressing this global issue, governance centres may be especially interested in the experiments of others as a key source of lessons (Aligica, 2014, p. 66; Goodin, 1996, p. 42; Hildén, Jordan, & Huitema, 2017).

An issue of course emerges with regard to the previous points about context. If context matters in policy-making, how can one learn from others? Following McConnell (2010), there are those who argue that policy is so contextual that nothing can be learnt across governance centres. By contrast, others contend that policies work irrespective of

the context through set mechanisms (e.g., the power of the market to efficiently allocate resources). Between these arguably extreme positions is what McConnell (2010, p. 200) terms the 'familial way' of contexts. In other words, while contexts may differ on a range of conditions, some settings are more similar than others. For example, if a country has a democratic parliamentary political system, all else being equal, a successful policy may be more likely to succeed in another country with a similar political system rather than a very different one (e.g., an authoritarian state). Thus, it may be possible to determine to what degree contexts are reasonably similar. This is of course no guarantee of success (McConnell, 2010). However, if a governance centre wishes to learn from the experiences of another, it may be helpful to decipher which contextual conditions were critical for the success of a particular intervention, and if those conditions are present elsewhere (see Benson & Jordan, 2011).

This view of automatic interactions driven by a range of incentives contrasts with insights from other governance literatures that point to the need to stimulate interaction in some circumstances (e.g., Jordan & Schout, 2006). There are reasons to believe that self-organization and consequently 'taking into account' may not be automatic, something which has stimulated numerous debates on 'meta-governance.' In the absence of strong market signals or other powerful incentives—which is often the case in the public sector where duplication of services may be seen as a waste of resources—other mechanisms may be necessary in order to generate enough pressure to compel governance centres to pay attention to one another. In other words, it may be necessary to externally induce some of the dissatisfaction that Rose (1991) considers essential for lesson-drawing to happen. This is also because while it may be perfectly rational from a collective standpoint to learn from others and continually improve governance practices, numerous factors such as vested interests, path-dependent behaviour, pre-existing institutions and general political inertia bolstered by overburdened policy-makers, may prevent such learning in practice, thus necessitating other forms of coordination. Hierarchies are one way to achieve this (see Peters, 1998), but hierarchy does not sit well with the Ostroms' ontology of self-organization and may in some cases may not even be possible (notably in the international climate regime at the time of writing).

In increasingly networked arrangements, where neither markets nor hierarchies force coordination, mutual taking-into-account, or what others have termed 'policy coordination', may thus be subject to substantial collective action dilemmas (Jordan & Schout, 2006), an issue that the whole polycentric governance approach seeks to address (E. Ostrom, 1990). Even though the system as a whole could benefit from learning, individual governance centres may not be able or willing to draw lessons from others or provide their own lessons. For these reasons, some higher-level incentives, if only through coordination, may be necessary to drive interaction in polycentric systems (see also Hale & Roger, 2013; Jordan & Schout, 2006). To make this happen, 'political pressure' or some resource provision from 'on high' may be needed (Jordan & Schout, 2006, p. 271).

Polycentric governance scholars have over time acknowledged the need for 'higher-level institutions' to some extent. In her work on polycentrism, Elinor Ostrom advocates a subtle blend of self-organization by local actors and "some larger-scale jurisdiction" (E. Ostrom, 2005, p. 282). Ostrom is less clear, however, on the origin and precise nature of this 'larger-scale jurisdiction.' On the one hand, she argues that sometimes pre-existing higher governance levels (e.g., state structures) are ineffective and it may therefore be advantageous to grow higher-level structures from lower levels:

> Success in starting small-scale initial institutions enables a group of individuals to build on the social capital thus created to solve larger problems with larger and more complex institutional arrangements (E. Ostrom, 1990, p. 190)

On the other hand, she argues on the same page that in a key case study on Californian water governance, recourse to pre-existing institutions such as the (public) court system proved vital in fostering self-organization among local actors. In a similar vein, Aligica (2014) stresses "[…] an over-arching system of rules […] (p. 57) as one of the 'three basic features' of polycentrism (p. 58)—which may be an 'institutional and cultural framework' (p. 58) that determines who participates in a polycentric governance system" (p. 59).

Based on the latter reasoning, Mansbridge (2014) argues that Elinor Ostrom in fact frequently alluded to higher-level governance functions that are often—but not always—conducted by states. Based on her reading of Ostrom, Mansbridge (2014) emphasizes that:

> Ostrom's polycentric model assumes some levels higher than the local, which can threaten to impose other solutions, provide neutral information, provide venues and support for the local negotiation, and, crucially, sanction non-compliance. (p. 9)

Mansbridge (2014) goes on to argue that more traditional public actors including states may deliver some or all of these four functions. Notably, although Mansbridge (2014) does not specifically define what she means by 'the state,' her discussion of fairly wide-ranging functions included in the above quote appears to allude to a broad definition of what precise institutions are thought to form part of the state. This is in line with a relatively broad description in the *Concise Oxford Dictionary of Politics,* which defines 'the state' as "[a] distinct set of political institutions whose specific concern is with the organization of domination, in the name of the common interest, within a delimited territory" (Burnham, 2009). Taken together, scholars working in the polycentric tradition would conceive of the state fairly broadly, including institutions forming the legislative, executive, and judicative branches. In sum, coordination or 'taking each other into account' may in some cases happen automatically, but in others require conscious effort and coordination. These questions have a direct bearing on the central questions of this thesis, namely where these 'lessons' are going to emerge from (i.e. who generates the lessons) and whether the lessons are provided in a way that can at least in principle enable lesson-drawing across governance centres (and thus the shift from polycentricity to polycentrism that Chapter 1 explains).

Crucially for this thesis, a focus on information provision and enforcement via monitoring is a central and explicit component in polycentric governance theory. As Elinor Ostrom (1999) explains,

> If all self-organized resource governance systems are totally independent and there is no communication among them, then each has to learn through its own trial-and-error process. (p. 525)

Some scholars highlight "[…] that a polycentric arrangement has a built-in mechanism of self-correction" (Aligica, 2014, p. 48) and advance the (big) claim that "[…] reflexivity is a systemic feature […]" (Aligica, 2014, p. 66). As Elinor Ostrom (1999) writes,

> Thus, a self-organized resource governance system with a higher level of in-migration or greater communication with other localities is more likely to adapt and change rules over time than is a system where new ideas concerning how to use rules as tools are rarely brought in. (p. 525)

But because reflexivity requires knowledge and critique of ongoing approaches, it depends on mechanisms to provide that knowledge. Otherwise, polycentrism, or "a system of reciprocal monitoring and assessment in dynamic interdependence" (Aligica, 2014, p. 66) may not materialize. But who will provide this information, will it appear with or without central stimulation, and will what emerges be of sufficient quality to be useful? Aligica was rather optimistic, assuming that

> A system of 'reciprocal monitoring and assessment for the range of institutions available in society' is thus put spontaneously in place, but in addition a system of broad checks and balances emerges. (Aligica, 2014, p. 66)

Others, such as Mansbridge (2014) envision a much stronger role for traditional public actors such as states, which could "*help* monitor compliance and sanction defection in the implementation phase" (p. 8, emphasis added). However, alternatively, states or other governance actors may shy away from the costs of collecting information about the experience in other governance centres or from making relevant changes once they know that another approach may generate better results. Thus, taken together, the question that runs through the literatures on common pool resources, polycentrism and policy coordination centres on who provides 'collective' or 'public' goods, which may

include the extent to which governance centres monitor their own practices and in turn pay attention to one another in order to learn and, perhaps, coordinate their activities.

2.4 Monitoring: from common-pool resources to climate policy

Common pool resource scholars in the polycentric governance tradition highlight that monitoring is an absolutely essential part of successful CPR governance. As Elinor Ostrom (1990, p. 45) emphasizes, "[w]ithout monitoring, there can be no credible commitment; without credible commitment, there is no reason to propose new rules." A fairly general definition holds that monitoring may be defined as "[a] continuing function that uses systematic collection of data on specified indicators to provide […] indications of the extent of progress and achievement of objectives and progress in the use of allocated funds" (OECD-DAC, 2002, p. 27-28). In other words, monitoring refers to "[…] recipe[s] for the selection, organization and retention of large amounts of information" (Dahler-Larsen, 2011, p. 65). Elinor Ostrom strongly links monitoring with the idea of preventing rule defections (i.e. policing).

But what makes monitoring particularly successful? Evidence from resource management literatures suggests that there is no general recipe for organizing monitoring activities. For example, Ostrom and Nagendra (2007) use multiple methods to show that the success of forest management depends critically on the fit of monitoring institutions with wider ecological, social, and political environments (or context, see above). Furthermore, the success of a monitoring regime often hinges on whether it is perceived as legitimate, which tends to be the case when people who are affected by the regime are involved in its creation and maintenance (E. Ostrom & Nagendra, 2007). Participants may then even be willing to bear some of the cost of monitoring themselves (E. Ostrom & Nagendra, 2007). The key lesson to take from these smaller-scale studies is that in some cases, decentralized monitoring appears to work 'better' than centralized activities for the reasons outlined above. But, again, the success of a particular monitoring regime depends critically on its fit with the particular context, including existing institutions, cultures and the nature of the resource. When monitoring is

successful, it can not only prevent rule defections, but also provide knowledge that may be of use to other governance centres—driven by self-organizing actors.

When moving to larger common-pool resources (such as the atmosphere and a stable climate), Elinor Ostrom argues that the more successful governance systems tend to organize "appropriation, provision, *monitoring*, enforcement, conflict resolution, and governance activities […] in *multiple layers of nested enterprise*" (E. Ostrom, 1990, p. 101, emphasis added). This is because "[e]stablishing rules at one level, without rules at the other levels, will produce an incomplete system that may not endure over the long run" (E. Ostrom, 1990, p. 102). Thus, monitoring by a single actor at a single level is unlikely to work in these instances.

At any level, monitoring is neither an easy nor a 'cheap' activity (E. Ostrom, 1990; see also Schoenefeld et al., 2018). Kusek and Rist (2005, p. 301) have noted that "[t]he reality is that putting in place even a rudimentary system of monitoring, evaluating, and reporting on government performance is not easy in the best of circumstances." Whether monitoring natural resource use or public policy, doing so requires significant and sustained effort, time, resources, and buy-in by multiple parties to set up and operate monitoring activities (E. Ostrom, 1990, p. 202). But not all monitoring activities are created equal. Importantly, Elinor Ostrom (1990) argues that for natural resources,

> Monitoring costs are affected by the physical attributes of the resource itself, the technology available for exclusion and appropriation, marketing arrangements, the proposed rules, and the legitimacy bestowed by external authorities on the results of institutional choices […]. (p. 203)

Furthermore, "[f]actors that enhance the capacity of users to see or hear one another as they are engaged in appropriation activities tend to lower monitoring and enforcement costs" (E. Ostrom, 1990, p. 204). Additionally, "[t]he availability of low-cost facilities for recording and disseminating information about regulated activities will also decrease monitoring costs" (E. Ostrom, 1990, p. 204). In other words, the more detectable an activity—and potential rule breaking—is, the easier it is to monitor.

The physical size of a resource also has a strong bearing on monitoring. Generally, "[t]he larger the resource, the greater the costs of 'fencing' and/or patrolling the boundaries to ensure that no outsider appropriates" (E. Ostrom, 1990, p. 203). And if frequent monitoring is required, costs tend to increase (E. Ostrom, 1990, p. 204). Finally, it is important to recognize that the nature of the rules to be monitored also affects the ease of monitoring:

> Rules that unambiguously state that some action – no matter who undertakes it – is proscribed are less costly to monitor than are rules that require more information about who is pursuing a particular behavior [sic] and why. (E. Ostrom, 1990, p. 204)

Furthermore, "[r]ules that place a limit on the quantity of resource units that can be produced during an entire season or year are more costly to enforce" (E. Ostrom, 1990, p. 205). The smaller and the more visible a resource and its use are, and the clearer the rules that govern it, the easier it is to monitor.

In cases where more technical scientific knowledge may be required to monitor a resource (such as overall fish stocks to determine fishing quotas), Elinor Ostrom points to the self-organizing capacities of local actors through community organizations. She argues that

> While no single community-governed organization may be able to fund information collection that is unbiased and of real value to the organization, a federation of such organizations may be able to amass the funds to do so. Simply having a newsletter that shares information about what has worked and why it has worked in some settings helps others learn from each other's trial-and-error methods." (E. Ostrom, 2005, p. 280).

Information generated in this way may be more sensitive to the interests and needs of the local actors who fund them—and help systemic learning. Crucially,

> Associations of local resource governance units can be encouraged to speed up the exchange of information about relevant local conditions and

> about policy experiments that have proved particularly successful. (E. Ostrom, 2005, p. 283)

Self-organization may in turn support interactions between governance centres.

There is thus a strong argument to consider the 'institutional fit' between what is being monitored and the institutions to do so. As Keohane and E. Ostrom (1994) explain:

> Another implication of research on local CPRs and public goods and on international regimes for international environmental institutions is the importance of *achieving a match between the characteristics of a successful monitoring and sanctioning scheme and the characteristics of specific situations.* (p. 22; emphasis added)

Table 2.1 summarizes the key insights from monitoring common pool resources with a view to applying them to monitoring climate policy in the next section in light of the three foundational ideas of polycentric governance theory identified above. For example, the nature of the resource relates to context, whereas information exchange through associations relates to interacting governance centres.

Table 2.1: Key insights from literatures on monitoring common-pool resources

Self-organization	• Actors have the capacity to self-monitor; doing so may increase legitimacy of a monitoring regime and ownership/buy-in. • If individuals or community organizations do not have the necessary resources to conduct (scientific) monitoring, they may form associations that pool resources.
Context	• The type of resource matters – some are much more difficult to monitor than others. • Larger systems are more difficult to monitor than smaller ones. • Clear-cut and precise rules are easier to monitor than more general ones. • It is important to consider the 'institutional fit' between a monitoring institution and its context (the resource, community structure, etc.).
Interaction	• Associations of organizations can stimulate the flow of information between governance centres; this can lead to learning from different experiments.

Sources: (E. Ostrom, 1990; E. Ostrom, 2005; E. Ostrom & Nagendra, 2007)

What can we glean from these insights on monitoring natural resources for monitoring climate change policy? A first thing to note is that humans cannot readily detect carbon dioxide and other greenhouse gases without significant technical equipment, making monitoring technically much more challenging than, say, monitoring the number of fish that have been taken out of a fishery. Monitoring greenhouse gases requires significant expertise and equipment, and has been subject to contestation, especially when there are direct policy consequences of monitoring decisions. For example, Canada and the EU have quarrelled intensely about the greenhouse gas content of tar sand oil (Neslen, 2011). It can also be extremely costly to accurately measure or estimate carbon emissions from certain sources—and may thus not viable in some cases (Öko-Institut, Cambridge Economics, AMEC, Harmelink Consulting, & TNO, 2012).

Second, the expanse of the atmosphere is vast and it is thus exceedingly difficult to establish boundaries for monitoring and 'appropriation.' Following Elinor Ostrom's rationale (see above), the physical nature of greenhouse gases makes monitoring rather challenging. It is hard to imagine how individuals may conduct such highly complex policy evaluations as they have been shown to do when monitoring individual resource governance.

But monitoring greenhouse gas fluxes is only one way of looking at policy outcomes, as other factors, such as impacts on congestion, public health, or employment are often equally at the centre of policy discussions—and the 'goals' of a policy may indeed be subject to significant contestation. Similarly, supply-oriented climate policy aims to leave a significant amount of hydrocarbons in the ground. Monitoring complex outcomes such as 'public health' typically requires the use of indicators, which "summarize or otherwise simplify relevant information, make […] visible or perceptible phenomena of interest, and quantify, measure, and communicate relevant information" (Gallopín, 1996, p. 108). However, using indicators to monitor policies is by no means a politically 'neutral' or 'innocent' activity, because these indicators embody underlying value orientations regarding what matters and what does not (Gudmundsson, 2003; Lehtonen, 2015) and are frequently constructed from information that is either readily available or can be generated (Gallopín, 1996). Even choosing indicators such as greenhouse gas emission reductions to compare climate policies embodies a deeply normative choice (Schoenefeld et al., 2018). The key difference is thus that CPR monitoring can often rely on direct measurement and observation of appropriation, whereas monitoring climate policies requires other tools to do so; and the goals of policy may be multifarious and sometimes fuzzy.

Furthermore, monitoring the effects of climate policies differs from that of common pool resources because policing and detecting rule defection is only one and possibly not the major objective of monitoring, which may also aim at learning, an aspect that features only partially in Elinor Ostrom's discussions of monitoring (see above). Related to the idea of indicators, climate policies may also generate a range of intended and unintended effects and potentially interact with other policies—as

discussed above, it is thus often necessary to use (multiple) indicators rather than direct observation; and it involves many more actors and jurisdictions. Last, because much may be at stake, climate policy monitoring tends to be so politically sensitive (Schoenefeld et al., 2018) that top-down monitoring has proven difficult if not impossible to do at the international level (see also Schoenefeld & Jordan, 2017).

In order to apply insights from CPR monitoring to climate policy monitoring, it is first necessary to somewhat relax the definition of 'local.' Clearly, the idea of local monitoring where one fisher(wo)man may observe the behaviour of her or his colleagues only has limited value when considering the monitoring of national climate policy. But if one allows the idea of more localized monitoring to apply to the nation state, it quickly becomes clear that some states and/or regions (and actors therein) do have the capacity to monitor their own climate policies. Thus, actors at 'more local' levels (here understood as national versus international) may be better placed—and viewed as more legitimate—to regulate their own actions. This view is certainly also in the spirit of the Paris Agreement, which relies on nation states putting forward their own contributions and assessing their progress over time (Schoenefeld et al., 2018). Similarly, when 'self-organization' is understood as an activity done at the nation state level (or certain actors within the nation state, which nevertheless do not necessarily have to be individuals), it becomes more feasible to apply these concepts.

Table 2.2 summarizes the key conclusions from the discussion in this section. Similar to what was done above, it organizes the points by the three foundational ideas, namely self-organization, context, and interacting governance centres.

Table 2.2: Monitoring (public) climate policy

Self-organization	• Self-monitoring can happen at national and sub-national levels (by both state and non-state actors). • Individuals/community organizations/states can pool resources to conduct monitoring.
Context	• Policy effects are difficult to monitor – many potential effects, greenhouse gases not easily detectable, lots of sources and actors. • The 'climate system' is very large (global). • It is difficult to define clear-cut and precise rules for monitoring, given technical issues and political sensitivities. • The 'institutional fit' between monitoring institutions and its context (the resource, community structure, etc.) matters for climate change, particularly when considering monitoring at 'lower' governance levels (national, regional, etc.).
Interaction	• Associations of organizations can stimulate the flow of information between governance centres; this can lead to learning from different experiments; this can also happen at the international level, e.g. EU – (see Schoenefeld et al., 2018).

Drawing on Table 2.2, scholars working in the polycentric governance tradition would thus likely ask with respect to climate policy monitoring: *how* do actors monitor climate policy, *what* do they include (or ignore), *who* conducts the monitoring, and how do those engaged in monitoring *interact* with one another and their context? These three core ideas relate closely to the foundational ideas of polycentrism, and thus become the basis for discussing the role of policy evaluation in the following section.

2.5 Evaluation in polycentric governance systems

Some form of knowledge generation on the effectiveness of policy approaches in different governance centres is part and parcel of polycentric governance. Empirical research on monitoring in common-pool resource systems (see above) contains necessary, but not yet sufficient insights to interrogate what role—if any—policy evaluation could and potentially already does play in polycentric governance systems. Compared with the definition of monitoring at the beginning of this section, *ex-post* policy evaluation is a related, and yet substantially different activity. Recall that this thesis follows Vedung (1997, p. 3), who defines policy evaluation as the "careful retrospective assessment of the merit, worth, and value of administration, output and outcome of government interventions, which is intended to play a role in future practical action situations" (see Chapter 1). Monitoring data may be an ingredient of evaluation, but evaluation goes a key step further than monitoring in making a value-based assessment, and evaluation can take a much broader view and consider factors and data that limited monitoring may struggle to pick up.

Policy evaluation is thus a broader activity than monitoring, and therefore, its role in polycentric governance system must also be considered in broader terms. There are two headline reasons why policy evaluation may, in principle, play a role in polycentric settings—and which are also frequently cited as reasons for evaluating to begin with (see Borrás & Højlund, 2015; Sanderson, 2002): first, related to Elinor Ostrom's ideas about detecting rule defection via monitoring, policy evaluation may play a role in enabling accountability relationships in polycentric systems (Versluis, van Keulen, & Stephenson, 2011, p. 206). Bovens (2007) defines accountability as "a relationship between an actor and a forum, in which the actor has an obligation to explain and to justify his or her conduct, the forum can pose questions and pass judgment, and the actor may face consequences" (p. 107). A key issue in 'new' governance contexts—including potentially polycentric governance—is that traditional forms of accountability, which are enacted through often long principal-agent chains, are becoming increasingly problematic (Stame, 2006). Whereas democratic states usually boast civil servants who answer to elected leaders who in turn answer to Parliament, which itself answers to its

voters, such a conceptualization of accountability struggles in polycentric settings, where it may be much less clear who answers to whom. This state of affairs has given rise to 'new forms of accountability,' such as diagonal or horizontal ones where policy-makers may be accountable to civil society or to ombudspersons (Bovens, 2007). From this perspective, policy evaluation may enable accountability in polycentric settings (Bovens, 2007). Alkin and Christie (2004, p. 12) have also highlighted that "[t]he need and desire for accountability presents a need for evaluation." Relatedly, Hanberger (2012) focuses on the role of evaluation to support political accountability in different governance systems, including state systems, regional-local systems and network governance. Policy evaluation may thus make a significant contribution to enabling accountability (Fischer, 2006). But more than providers of 'objective' policy information, evaluators may also be seen as mediators between societal discourses and discussion about the merit of particular policies to achieve a number of different—and layered—goals (Fischer, 2006). In this model, evaluators are not aloof from society, but inextricably bound up and working within and through a system of values and facts that are at stake when a policy is evaluated.

The second, and certainly no less widely discussed reason why policy evaluation may have a role to play in polycentric governance systems is as an enabler of learning (see Section 2.2 above for a definition and discussion of the concept). Scholars have already highlighted potential links between *ex-post* evaluation and learning. For example, Haug (2015, p. 5) stresses that "[e]x-post evaluation of programmes or policies […] is a widely applied group of approaches aimed at stimulating learning in environmental governance." There is still an ongoing and largely unresolved debate on what exactly is learned, which depending on one's philosophical position may range from 'facts' to learning about value-based discourses (Borrás & Højlund, 2015; Haug, 2015; Sanderson, 2002). This thesis focuses on the learning-related factors that feature most strongly in debates on polycentric governance, namely the importance of context, as well as learning as one vehicle of interactions between governance centres (see above).

While accountability and learning may be two theoretically relevant concepts for understanding a potential role of policy evaluation in polycentric governance systems, it is important to recognize that policy evaluation happens in a political environment and may therefore also be done for political - that is strategic and 'irrational' - reasons that have little to do with either accountability or learning. 'Political' in this context is understood as both processes and struggles that happen inside familiar governmental arenas, but also as a more pervasive process that happens when power operates, and regarding what is discussed and addressed in public and what is not (Hay, 2002; Lukes, 2005; Mansbridge, 1999, p. 214). Numerous scholars have already highlighted the political characteristics of policy evaluation (Bovens et al., 2006; Greene, 1997; House & Howe, 1999; see also Lascoumes & Le Gales, 2007; Nilsson et al., 2008; Owens, Rayner, & Bina, 2004; Vedung, 1997). First, Weiss (1993) argues that because government programmes emerge through political processes, political pressures are unlikely to disappear at the evaluation stage (though they could arguably change over time). Second,

> As social scientists increasingly recognize, no study collects neutral "facts": all research entails value decisions and to some degree reflects the researcher's selections, assumptions, and interpretations. (Weiss, 1993, p. 102)

Third, policy evaluation may also be political because it has the potential to affect the range of decisions political actors can take and thus act as a 'destabilizing' force (which links with the points on 'dissatisfaction' with regard to lesson drawing above). For example, a supportive evaluation may provide vital support to continue or extend a climate policy, whereas a negative evaluation may deprive decision-makers of the possibility to do so and can potentially lead to policy dismantling (see Gravey, 2016). Fourth, because policy evaluation has the potential to affect resource distribution across society (Bovens et al., 2006), it may be used in a strategic fashion such as to delay a political process or to move a decision to another forum. Thus, policy evaluation is political because it operates in a political context, can destabilize resource distribution, and can be used in a strategic way.

These political, and often strategic and normative elements of evaluation generate crucial but difficult questions for the role of policy evaluation in polycentric governance settings. If evaluation is done for more strategic and political reasons (see also Pollitt, 1998), then the outcome of evaluation may be less-than-optimal from a polycentric governance perspective, and thus expectations towards evaluation may have to be tempered. By the same token, in situations of considerable political contestation, it is also possible that evaluative knowledge may emerge through self-organizing capacities by individual governance actors (see below). For example, informal actors may conduct or commission their 'own' evaluations in order to contest points made by formal (i.e. state) evaluations. A whole range of evaluations may therefore generate a more 'complete' body of evaluative knowledge that does not rely on a single perspective. Thus, a polycentric governance perspective on evaluation would highlight the need for a broad range of evaluation perspectives and actors so as to generate diverse knowledge of policy effects.

Against this important background, the remainder of this section reviews debates in existing literature on policy evaluation insofar as they relate to the three foundational ideas of polycentrism, namely self-organization, context, and interaction between governance centres. Where pertinent, the review connects with ideas on accountability and learning, while keeping in mind the political nature of policy evaluation.

2.5.1 Self-organization

This section draws on multiple strands of argument that have emerged from wider discussions on the role of actors in evaluation in order to develop insights into the role of self-governance in policy evaluation in polycentric settings. An understanding of what we currently know about who conducts, participates in and benefits from evaluation is crucial to theorizing the role of evaluation in polycentric settings. In order to map the literature, the section draws on numerous conceptual categories that have emerged in evaluation literatures over time. These include (1) who conducts evaluation,

including ‘contracted’ evaluation, formal and informal actors and the role of participation; and (2) who are the intended ‘users’ of evaluation.

Multiple actors may in principle be capable of evaluating policy (see Ostrom, 2005). A key point from the earlier discussion of common pool resource monitoring is that it matters a great deal who evaluates, for what purpose, and funded by whom. For analytical purposes, evaluation literatures have found it useful to distinguish between *formal* and *informal* evaluation. In an early article, (Weiss, 1993) distinguished between ‘inside evaluation’, which is conducted by people inside government, and ‘independent evaluation’ by people not linked with government (see also Chelimsky, 2009). Weiss (1993) argues that the uptake of ‘inside evaluations’ may be higher because in-house evaluators may have a better understanding of the policy-making environment, but that the findings are also likely to be less radical. By contrast, ‘independent evaluations’ are thought to take a much more critical look at policies. Other researchers have recently developed a related notion of ‘formal’ versus ‘informal’ evaluation, particularly in the climate policy sector in the European Union (Hildén et al., 2014; Huitema et al., 2011). Hildén, Jordan, and Rayner (2014) define formal evaluation as ‘state-led’ and informal evaluation as “evaluation activities by non-state actors” (p. 885). In sum, there are numerous actors who can become involved in evaluation activities, but knowledge on the impact of different actors on policy evaluation is only just emerging (for a review, see Schoenefeld & Jordan, 2017).

Relatedly, there are different views about where evaluation originates from. For example, Sager, Widmer, and Balthasar (2017, p. 316) have argued that “evaluation is not or mainly not self-motivated like basic research, but rather requires a demand in the form of commissioning actors” (translation by the author). By contrast, Elinor Ostrom (2005) has pointed to the potentially self-organizing capacities in scientific assessment, monitoring and potentially policy evaluation (see above). The available evidence thus far suggests that particularly governmental actors frequently commission evaluations. Pollitt (1998) has highlighted that in Europe, governments are among the most important evaluation sponsors. For example, a survey of climate policy evaluation in the EU showed that nearly half of all climate policy evaluations were commissioned

(Huitema et al., 2011); the rest were funded and conducted by the same organization. However, it should be noted that a footnote explains that many of the non-commissioned evaluations may have emerged from academic research projects, as this particular study used a wider operational definition of policy evaluation than that applied in this thesis (see Chapter 4). Differing definitions may thus also be one reason why scholars arrive at different conclusions regarding the self-organizing capacity of policy evaluation actors.

While in an ideal world, commissioning would add an extra dose of independence to evaluation (see Chelimsky, 2009), emerging research suggests that in practice, it can be the site of political struggles where those who commission evaluations often try to control their contractors (Pleger & Sager, 2016). For example, a survey of evaluators revealed that governments may seek to directly influence commissioned evaluators (Hayward et al., 2013) or at least frame evaluation findings in a more positive light (Weiss, 1993). According to Hayward et al. (2013), governments have a range of strategies to do so – for example, by controlling the research questions in an evaluation, or by enacting budgetary-turned-methodological constraints—e.g., not enough funding for a control group (see Pleger & Sager, 2016 for a systematic approach). Thus, in contracted evaluation, the emerging principal-agent relationships have at least the potential to be fraught with politics. Those who instigate an evaluation may not necessarily conduct it or intend to use it (Pleger & Sager, 2016), although of course all three activities can—at least in principle—be done within a single institution or even by a single person. The aforementioned distinction between 'formal' and 'informal' evaluation becomes significantly more difficult once multiple actors become involved in a single evaluation (see Chapter 4).

While evaluation literatures have long problematized the relationship between formal and informal evaluation and their influence on evaluation results, early scholars often considered formal and informal categories rather crudely i.e. paying insufficient attention to principal-agent relationships between evaluation funders and evaluators whenever evaluations are commissioned (Weiss, 1993). Emerging evidence challenges this limited view by suggesting that the process of commissioning evaluations correlates with evaluation results: Huitema et al. (2011), for example, show that climate policy

evaluations that were commissioned are much less reflexive (i.e. critical of extant policy targets) than evaluations that were not commissioned. There is thus an urgent need to further explore the influence on evaluation outcomes when both formal and informal evaluators commission evaluations. For example, Hayward et al. (2013) consider this principal-agent relationship and show how (formal) evaluation funders (British civil servants) sought to influence evaluators at various points.

With a view to climate policy, some earlier scholars have made strong prescriptive statements on who 'should perform' climate policy evaluation. For example, Feldman and Wilt (1996) argue that informal (i.e. non-state) actors have a particularly critical role to play because "[…] evaluation of these [climate change] programs must ultimately be performed by some external entity, group, or institution" (p. 67). They go on to argue that

> Whereas NGOs [non-governmental organization] may certainly have their own agendas, as a supplement to national and international organization review of subnational plans, NGO review may provide alternative data, complementary criteria for evaluation, or other important information that could help improve the evaluation, and thus performance, of national climate action plans (Feldman & Wilt, 1996, p. 67).

However, in line with Elinor Ostrom (2005), Feldman and Wilt (1996, p. 66) also suggest that "[…] national-level guidance, particularly in commissioning research, is needed to ensure data quality." Thus, these authors assume the need for a higher-level jurisdiction in assisting the evaluation of climate change policy by informal actors.

A second way to look at self-organization is through public participation in evaluation. In general, prescriptive evaluation literatures have over time widened the circle of contributors to evaluation. For example, Vedung (2013) explains three evaluation models based around the actors that evaluation seeks to involve. For example, in the 'client-oriented model', clients, or the 'receivers' of policy, evaluate the policy according to their own criteria. There has certainly been no shortage of additional

approaches in the prescriptive tradition to encourage greater participation of actors with a 'stake' in evaluation. For example, the 'empowerment evaluation' approach aims at 'empowering' those with a stake to participate in evaluation, while the evaluator is seen as a moderator who generates the circumstances in which people can empower themselves (Fetterman & Wandersman, 2005). The approach has devout followers – for example, Diaz-Puente, Yaguee, and Afonso (2008) describe how they used empowerment evaluation in Spain to evaluate projects with EU structural funding in the Madrid region. However, the fact that the authors were also the evaluators, their overly positive assessment of the method, their claim that it is perfectly compatible with EU evaluation requirements, and their use of only positive quotes from participants in this evaluation leaves some doubt regarding the potential critical voices that may have been omitted in this particular article (Diaz-Puente et al., 2008). But not all participation is equally 'empowering.' For example, individuals may simply be asked how satisfied they are with a particular service. Other approaches in the participatory tradition go farther to suggest that those affected by a policy should participate directly in evaluation and that evaluators hence become facilitators of an emerging dialogue between various individuals (Fischer, 2006). Some evaluation methods (e.g. surveys or interviews) are much more participatory than others such as formalized modelling. Thus, one way to assess the level of public participation in policy evaluation is to look closely at the evaluation method and set-up.

Another way to distinguish between more or less self-organizing evaluations is to consider whether or not they respond to a legal requirement to evaluate, often in the form of an 'evaluation clause' in legislation. There are, in principle, different types of evaluation clauses, ranging from general ones to clauses that apply to the activity of specific institutions or areas of administration (Bussmann, 2005). Emerging evidence suggests that legislation now commonly includes legal requirements to monitor or evaluate policy outcomes at regular intervals. For example, Mastenbroek et al. (2016) found that out of the 216 European Commission *ex-post* legislative evaluations they identified, 81% responded to an evaluation clause. In another case, Bundi (2016) explains that Switzerland introduced a general evaluation clause in its constitution in 1999. By 2008, Bussmann had identified about 90 such clauses at the national level in

Switzerland. Evaluation scholars have taken the increasing presence of evaluation clauses as an indication of advanced evaluation institutionalization (Jacob et al., 2015). This thesis uses the presence of evaluation clauses as a way to indicate the level of 'self-organization' – an evaluation that responds to a legal requirement can be considered one of the least self-organized. However, there appears to be little data on the existence of evaluation clauses or corresponding evaluations in the climate change sector.

In sum, there are numerous questions that emerge from this review. Although the formal-informal distinction has proven a useful conceptual tool, it remains an open question to what extent the categories of 'formal' and 'informal' evaluators blur or even interact, as has been suggested in other policy areas (Guha-Khasnobis, Kanbur, & Ostrom, 2006). Furthermore, the above discussion explains how thinking about policy evaluation in the polycentric governance tradition would not stop at simply adding more actors or methodologies. This view would crucially pay attention to how these actors interact in their evaluation endeavour. The following section focuses on this core issue.

2.5.2 Context

The idea that context matters in policy evaluation is not new, but is certainly contested. The *Encyclopaedia of Evaluation* defines context as "the setting within which the evaluand (the program, policy, or product being evaluated) and thus the evaluation are situated. Context is the site, location, environment, or milieu for a given evaluand" (Greene, 2005, p. 83). The entry then goes on to emphasize that context "is an enormously complex phenomenon" (Greene, 2005, p. 83). Other evaluation scholars have echoed these arguments. For example, Vedung (1997, p. 213) explains "that explanations involving administrative action are circumstantial. Universal explanations, valid for all times and regardless of surroundings, simply do not and cannot exist in the social world." Theorists proposing 'realistic evaluation' have argued that mechanisms (i.e. the connection between cause and effect) operate within contexts, and evaluators need to pay close attention to both the former and the latter in their endeavours (Pawson & Tilley, 1997, p. 63-78). More fundamentally, Guba and Lincoln (1981, p. 39-47)

argue that the merit and worth of a policy depends critically on the context; policies that may be valuable in one context could exhibit little value in another. Taken together, Patton (2008, p. 40) stresses that

> Program evaluation is undertaken to inform decisions, clarify options, identify improvements, and provide information about programs and policies *within contextual boundaries of time, place, values and politics*" (emphasis added).

As Tilly and Goodin (2006) argue in their introduction to the *Oxford Handbook of Contextual Political Analysis*, these are impressions of a more long-standing debate between those who hold that political processes have general attributes that are stable over contexts and time, and those who argue that political outcomes are highly contingent with regard to context (see also Pollitt, 2013). While some argue that there are mechanisms that function independently of contexts, others such as Martin (2001, p. 204) highlight that "local context matters in the formation and practice of policy" and Kaufmann and Wangler (2014) add that this holds especially in the case of environment and climate policy. In the area of evaluation, Guba and Lincoln (1989, p. 45) have for example argued that "[p]henomena can be understood only within the context in which they are studied; findings from one context cannot be generalized to another; neither problems nor their solutions can be generalized from one setting to another." But others, such as Pawson and Tilley (1997, p. 22) disagree in arguing that generalizations of context-bound mechanisms may indeed be possible. In practice, both elements are likely to emerge—for example, the EU greenhouse gas emissions trading scheme drew on experiences with sulphur dioxide trading in the United States in a more or less instrumental way. However, following the experiences in the EU, actors such as California and Australia were able to gain a much richer, contextual understanding of the struggles that emerged with this instrument (particularly the impact of the global financial and economic crisis) and design their own instruments accordingly (Bang, Victor, & Andresen, 2017; The Economist, 2014). Thus, evaluations that seek to 'correct for context' by making contextual variables explicit, but that still seek to

identify some general 'lessons' may prove most adequate in polycentric settings (see Tilly & Goodin, 2006). As Greene (2005, p. 84) asserts,

> All evaluators agree that context matters, for the programs, policies, and products we evaluate and for the conduct and probable effectiveness of our work as evaluators. All evaluators also agree that good evaluation is responsive to, respectful of, and tailored to its contexts in important ways.

Such arguments have also been advanced in more scholarly debates. For example, Wells (2007) argues that "evaluative research undertaken with an understanding of political ideas, institutions and contexts provides a richer basis on which to inform policy, and equally, practice" (p. 27). Overall, Fitzpatrick (2012) notes in her review of the evaluation literature that attention to context has continuously featured in writings on evaluation since the early days in the 1960s and 1970s; yet, she also writes that "context is an amorphous issue" (p. 7). Polycentric governance scholars, too, would strongly reject the argument that public policy generates comparable effects regardless of contexts, making direct, instrumental learning challenging. By contrast, they would emphasize that because contextual factors generate highly idiosyncratic pathways of policy development, direct, instrumental learning may be difficult—though other forms of learning, such as political learning, which involves gaining knowledge of the political preferences of others or drawing lessons in context may still take place (see Zito & Schout, 2009 for a discussion of different types of learning). Given the clear arguments on context by polycentric governance scholars (see above) this thesis works in the latter tradition.

There are generally two ways in which evaluation literatures propose to deal with context. The first includes accounting for contextual factors either in an inductive or deductive way, and scholars have started cataloguing potential factors that may matter, while emphasizing differences across policy areas. This section begins with a general discussion of potentially relevant contextual factors and then turns to factors that are especially discussed in literature on environment and climate policy evaluation. A second way in which context may be dealt with in policy evaluations is through the

evaluation approach, for example expressed through the evaluation methodologies used or the criteria applied. The second part of this section thus turns to the relevant discussions in this area.

On accounting for contextual factors in policy evaluation, Greene (2005, p. 84), who has a social psychology background, highlights contextual dimensions such as demography, material and economic aspects, institutions and organizations, personal interactions and norms, as well as politics as key contextual factors. Seven years later, Rog (2012) proposes a new framework, which identifies five key areas where contextual factors could be considered in policy evaluations: the nature of 'the phenomenon and the problem' (e.g., how much is known about the problem); the 'nature of the intervention' (e.g., how complex it is), and thus the need for multiple indicators and multiple methods and pathways to understanding effects; the 'broader environment/setting' including potentially layers of administration or institutions; 'the evaluation context' such as the budget or time available for evaluation; 'the decision-making context,' including the evaluation audience and its needs. In each dimension, there are "physical, organizational, social, cultural, tradition, political and historical" elements to consider (Rog, 2012, p. 27). However, the conclusion of the special issue stresses that this framework should not be applied in a 'rigid' manner; in fact, assessing context still requires 'subjective judgements' and skilled evaluators, given the plethora of potential contextual effects (Conner, Fitzpatrick, & Rog, 2012).

Based on such earlier work, Vo and Christie (2015) reviewed relevant literature and proposed an even broader framework in order to consider context in evaluation, namely one that focuses on the "who, what, where, when, why, and how (including "how much," which deals with valuing and is unique to evaluation)" (p. 48-49). The core argument here is that the contextual factors that other studies have catalogued (see Greene, 2005; Rog, 2012) proved too specific. However, given the specific focus of this thesis on climate policy evaluation, it is still useful to identify potential contextual factors within the specific field of climate policy evaluation. What, then, are the contextual factors that have already been discussed as particularly relevant for climate policy evaluation? The paragraphs that follow review the factors of time, geography and

spatial aspects, policy effects, external shocks and influences, and the political environment and structures. While this is clearly not an exhaustive list, these factors provide starting points that have received considerable scholarly attention in the past and which are likely to be relevant for climate change policies.

Time: While there may have been a time when scholars considered policy-making largely a-temporal and independent of the effects of time, more recent discussions in public policy and management have sought to re-introduce the variable of time (see Pollitt, 2008). These general debates have also been addressed in the context of policy evaluation. For example, Bressers, Twist, and Heuvelhof (2013) argue that time introduces a key element of complexity and unpredictability into public policy. This is especially relevant for environment and climate policy, which often exhibits 'time-lag effects' (Crabbé & Leroy, 2008, p. 38). For example, a policy that changes fundamental aspects of energy infrastructure may take a significant amount of time to take effect and produce measurable outcomes, given significant lock-ins in the sector (a power plant may take several decades to recover initial investments and produce returns, for example). Further, particularly with regard to climate change, the on-the-ground effects of a policy may play out over very long time scales (Mickwitz, 2003). Importantly, effects may develop over time, and short-term positive effects may not necessarily translate into long-term policy success (McConnell, 2010, p. 92). For example, in climate policy, a shift from coal to natural gas generates short-term reductions in greenhouse gas emissions because natural gas produces less carbon dioxide per unit of energy than coal, but locks the energy infrastructure into using fossil fuels for decades to come (unless there are viable alternatives to natural gas). Therefore, scholars generally recommend evaluating policies over time (the longer the time scale the better), and considering a wide variety of intended and unintended effects (Bressers et al., 2013; Kaufmann & Wangler, 2014; Mickwitz, 2013; Mickwitz, 2003). From the perspective of addressing climate change, long-term success ultimately matters much more than short-term effects that may prove transient and or even counter-productive. Taking into account a longer time horizon also matters because "policies rarely have a fixed beginning and end; usually new policies are piled upon old ones, or policy goalposts are shifted" (Crabbé & Leroy, 2008, p. 39). Thus, a short time horizon may miss crucial

elements in policy development and effects. In a similar vein, Hildén (2009) argues that taking into account a longer time horizon allows identifying path dependencies and outcomes that may have nothing to do with the policy intervention. Vedung (1997) further argues that legislative history may affect the outcomes of policy interventions, driven for example by the strength of political support at the time of instituting an intervention or the participation of affected parties in the policy-making process (p. 213-219). Taken together, evaluation theorists thus suggest expanding *ex-post* evaluation from a snapshot to a more long-range view, which potentially includes even the time before an intervention started.

Geography and spatial aspects: There are two key issues of importance here: one is the physical geography of a jurisdiction where a policy is implemented. Offshore wind energy, for example, may be an effective policy choice for the UK precisely because the country has ample coasts with comparatively shallow waters where erecting wind turbines is a viable option. By the same token, Norway may be particularly well-suited to hydro power, whereas southern Spain has geographical conditions that are particularly suited to solar power. Taking such factors into account will likely improve the understanding of policy effects and be a key element in lesson-drawing.

The second issue is a broader, spatial consideration that is ultimately tightly linked with what concerns polycentric governance: policy outcomes may to a great extent depend on the characteristics of the governance centre where they are implemented. As Crabbé and Leroy (2008) remind us, environmental issues often cross borders, and policies are often most effective when they address the scale at which the problem is caused (p. 39). While there are various conundrums about the causes and consequences of climate change in a broader sense (e.g., historical and current distributions of greenhouse gas emissions, as well as climate impacts) that have been discussed elsewhere (e.g., Raupach et al., 2014), the key issue here is the extent to which a policy has been applied at the 'right' scale. Arguably, putting in place an emissions trading policy versus planning local bike infrastructure is probably done best at different governance levels. Thus, evaluators may pay attention to scale in their evaluation. But the logic also goes the other way around: given their contextual nature, policy impacts

may not be evenly distributed across space, and success or failure may very much depend on that distribution (Martin, 2001). In sum, in order to understand the impact of geography on policy outcomes, it is relevant to understand whether evaluators discuss and analyse these dimensions. Thus, paying close attention to the physical, but also the socio-political factors that play a role in generating policy outcomes should be part of policy evaluation (Martin, 2001).

Policy effects: Given the highly complex nature of environmental policy systems (Crabbé & Leroy, 2008; Mickwitz, 2013) and potential emergent effects, policy evaluation scholars have argued that it is necessary to go well beyond the 'official' policy goals defined at the outset, and rather consider a range of policy effects, including unintended ones (Kaufmann & Wangler, 2014). Thus, the argument goes that it is necessary to consider a wide range of evaluation criteria in order to capture both intended and unintended main and side effects. Crucially, these effects also include interactions with other policies (Kaufmann & Wangler, 2014), given that policies hardly ever produce effects in isolation (Crabbé & Leroy, 2008, p. 39). Sometimes, a policy may be effective precisely because it is functioning in unison with others (such as siting policies to support subsidies for wind turbines). However, at other times, policies may detract from each other or be in conflict, such as providing subsidies for renewable and fossil-fuel based energy production (see Sorrell et al., 2003). Taken together, policy makers should thus consider a wide range of policy effects, as well as causal explanations that extend well beyond the logic of a singular policy.

Going beyond original policy goals has also been described in terms of reflexivity, especially with a view to climate policy evaluation (Fischer, 2006; Huitema et al., 2011). 'Reflexivity' in evaluation may be understood as the willingness to challenge extant policy goals (Fischer, 2006; Huitema et al., 2011). Given both the aforementioned 'political' context of evaluation, it is important to recognize that this context may include ill-defined policy goals, and that the entire context may shift over time (see also Conner et al., 2012). Thus, scholars have argued for more reflexive policy evaluations, or the idea that evaluators critically examine and if applicable revise the extant policy goals set at the initiation of policy.

External shocks and influences: External events, whether they are natural disasters, economic developments or other large-scale shocks can at times fundamentally change the overall system in which a policy operates. As Vedung (1997, p. 224) explains,

> The larger environment impacts on the outcomes. A program may be inherently clear, perfectly communicated to implementers, meticulously executed according to plan, and yet basically ineffective because of changes in the larger policy environment that upset the initial prerequisites for implementation.

For example, the global recession that began with the financial crisis in 2008 has arguably contributed significantly to (unexpectedly) reaching emissions reductions goals in Europe because of lower overall economic activity (Jacobs, 2012). Indeed, in this example, European climate policies may have contributed little – or not at all – to the achievement of that goal (see Kerr, 2007). In other circumstances, wider economic shifts, such as shutting down decrepit industries in East Germany after reunification in 1990 or the 'dash for gas' in the UK can generate significant greenhouse gas emissions reductions in the absence of an explicit intention to do so through climate policy (Jordan et al., 2010). Greenhouse gas emissions may decrease as part of a regular industrial transition towards a more diverse and service-based economy. Thus, where applicable, evaluators need to consider such external developments in order to generate a fuller understanding of policy impact.

Political environment and structures: General factors of the political environment, at times based around the way in which an intervention came about in the first place (see above), and at other times based around implementation, can influence the success of a policy and are thus crucial knowledge when seeking to understand the effectiveness of an intervention (Weiss, 1993). Vedung (1997, p. 226 – 245), for example, draws on implementation theory to explain how the nature of implementers, and especially their comprehension, capability and willingness to implement has an important bearing on outcomes. For example, a government agency that understands the intervention, has the necessary capabilities (e.g., financial resources, personnel and

equipment) and the willingness to implement is much more likely to implement successfully than an agency where the opposite is true. By the same token, the nature and reaction of the receivers of an intervention influences outcomes (Vedung, 1997, p. 238 – 241). For example, if a government implements subsidies for renewable energies, there is likely to be more uptake among a population that is well-informed about the existence of the intervention, and that has the necessary resources to make investments in order to capture these subsidies than among a population where the opposite is true. Finally, as Vedung (1997, p. 241 – 245) explains, policy outcomes also likely depend on interactions with other policies (sometimes strengthening, sometimes detracting from the policy – see above), as well as wider networks of stakeholders in support or in opposition to an intervention or the role of the media. All these factors related to the wider political environment have a potentially important role on the outcome of an intervention.

The second approach to consider context in evaluation is through conscious choices in the evaluation approaches, including the evaluation methods and criteria. With a view to the dimensions of her framework (see above), Rog (2012, p. 27) proposes using several methodological approaches, notably including stakeholders in the evaluation; using multiple methods; using quantitative indicators and explaining their variation. She stresses that "[h]ow we measure and incorporate context measures in each evaluation will likely have various levels and focus on relevant aspects of the each area of context (political, cultural, social, organizational)" (Rog, 2012, p. 37). The argument to use multiple methods has also been advanced by other evaluation scholars: Fischer (2006) lists key methodologies including 'experimental program research', 'quasi-experimental evaluation', 'cost-benefit analysis', and 'risk-benefit analysis.' But even when one uses particular models, Elinor Ostrom highlights that "[m]odels are useful in policy analysis when they are well-tailored to the particular problem at hand. Models are used inappropriately when applied to the study of problematic situations that do not closely fit the assumptions of the model […]" (E. Ostrom, 2005, p. 29). Thus, analogous to the 'institutional fit' in monitoring, Elinor Ostrom's arguments can be extended to consider the 'methodological fit' of monitoring as well (and, by inference, tailoring

methodologies to contexts). Prominent evaluation scholars have echoed this argument: As Toulemonde (2000, p. 356) writes,

> I consider it a universal rule that a good evaluation is "custom made"; in other words, each evaluation is unique […]. A good evaluation is designed at a given time, for specific users and in a specific context.

These insights may also hold for other evaluation methods, in that interviews and surveys can be adjusted to a particular policy and its context. In order to capture the full range and particularly higher levels of analysis, Fischer (2006) argues that qualitative methods such as interviews, participant observation and stakeholder surveys are particularly useful to 'get inside the situation'—or the context. For climate policy evaluation, the Öko-Institut et al. (2012, p. iv) emphasize that there is no "one-size-fits-all solution," and in some cases context may matter more than in others.

Very similar arguments on multiple methods have also been advanced in the realm of environmental policy. Mickwitz (2003) emphasizes in his framework for environmental policy evaluation that the complex nature of many environmental issues, and their uneven and at times remote effects make for an especially challenging treatment of context (see also Rog, 2012). He thus recommends using multiple methods, multiple criteria, as well as side-effect evaluation, intervention theories and participatory aspects in order to understand the multifarious effects of environmental policy in context (Mickwitz, 2003). Thus, a polycentric approach would advocate multiple and, in the best case, 'tailored' methods in policy evaluation.

Related to the idea of multiple evaluation methods is a debate that deals with using multiple evaluation criteria (Majone, 1989). Policy evaluation scholars and practitioners have emphasized the need to substantially widen policy evaluation criteria than what has been done earlier. As Vedung (2013) explains, "[i]n earlier literature, public sector evaluation *was* goal-attainment appraisal, period" (p. 389, emphasis in original). Using the goal-attainment approach, evaluation seeks to understand to what extent and how a particular public policy reached its own, predefined policy goals. However, the realization that goals may be ill-defined and that policy may generate significant

unforeseen effects due to contextual factors became a driver to conduct 'side-effect evaluation' that pays attention to a much wider range of policy impacts (Vedung, 2013). Knowing about wider and at times unpredictable policy effects led evaluators in turn to develop the 'relevance model,' where evaluation asks to what extent policy solves the 'underlying problem' that it seeks to address, even though policy impacts may not be in line with earlier predictions (Vedung, 2013). Fischer's (2006) key book also advocates using a broader spectrum of evaluation criteria, ranging from program verification (often described as goal attainment elsewhere) to situational validation (is the particular policy relevant to the situation its seeks to address?), societal vindication (does the program provide value for society as a whole?); and finally social choice (do the values that are behind the policy provide a good way of solving conflict?). It is thus clear that prescriptive evaluation theory has widened its criteria over time, and that this was done, at least implicitly, with a view to the importance of context in evaluation.

The idea of broader evaluation criteria also chimes with recent theoretical developments in polycentric governance theory. As Aligica (2014, p. 1) explains,

> […] when it comes to organizing human coordination and interdependence in diverse circumstances, with diverse preferences, endowments, and beliefs, institutional pluralism is a fact, a challenge, and a *prima facie* normative answer. If that is the case, then **the pluralism of criteria and values should as well define the way institutions and their performance are assessed** (bold emphasis added).

Using a small set of singular evaluation criteria will unlikely do justice to the contextual richness of many (polycentric) governance arrangements.

Taken together, contextual factors related to history/time, geography, intended and unintended policy effects, external shocks and influences and the general political environment are potentially relevant factors in climate policy evaluation. However, true to ideas about the contextual nature of policy, it would be difficult if not impossible to create an exhaustive, *a priori* list of factors that are likely to matter for climate change policy in particular. The above tentative list should thus be understood as a starting point

for the empirical investigation (see the following chapters), rather than as a definite statement. It should also be noted that

> Not all interventions are as susceptible to their contexts and not all investigations have to study each area of context with the same level of rigor and intensity used to study the core elements of a program and the outcomes (Rog, 2012, p. 37).

The above section has shown that there are numerous ways in which climate policy evaluation may pay attention to context, ranging from individual contextual dimensions to methodological adjustments. For example, context-sensitive evaluation may be able to shed a light on important co-benefits at varying scales in addition to reducing global carbon dioxide emissions (E. Ostrom, 2010b, p. 553; Somanathan et al., 2014). Crucially, paying attention to context also matters to the two headline concepts: for accountability, context-conscious evaluation can be a way to account for the whole range of policy effects, both intended and unintended. For learning, contextual information can provide crucial knowledge on a range of contextual mechanisms that brought about policy effects.

2.5.3 Interaction

As noted above, one of the key (normative) aspects in moving from polycentricity to polycentrism is that independent governance centres take each other into account and, ideally, learn from each other. Learning from each other necessitates some mechanism through which governance centres can know what happens elsewhere and bear in mind the contextual aspects discussed earlier. As reviewed above, there is some recognition in polycentric governance literatures that decentralized approaches may only be able to generate limited scientific information, particularly when dealing with larger governance systems (E. Ostrom, 2005). However, when such information becomes available vis-à-vis policy evaluation, it may be useful to foster the learning processes that polycentric governance scholars envision. In principle, policy evaluation could play a key role in facilitating this ‘taking into account’ through making activities in multiple centres

visible and intelligible. This is particularly relevant, because in order to benefit from governance experimentation in polycentric settings, "[…] we ought, furthermore, to encourage reflection upon the lessons from elsewhere and a willingness to borrow those lessons where appropriate" (Goodin, 1996, p. 42). For example, writing on the role of policy evaluation in the EU, Stame (2006) highlights that

> Just because the national states and the regions are so different, and thanks to the fact that public, private and civil society actors are neither absent nor mute, *there would be a great scope for listening to what the local situations have to say, scope also to compare the working of mechanisms in different contexts,* for creating a new body of European knowledge […]. (p. 14; emphasis added)

This remains a rare example, however, as in the past evaluation scholars have seldom considered such interactions across governance centres. Various factors may make lesson-learning across governance centres more or less likely. A crucial first step is that policy evaluations must become available to other governance centres in order to be able to have an effect. When governance actors can easily obtain evaluations from other governance centres (for example through indexed databases), they may be in a better position to use them (see Schoenefeld & Jordan, 2017). Once this is the case, the nature of the evaluations also matters. For example, executive summaries can add to the clarity of evaluation reports and help (busy) policy-makers to quickly assess whether an evaluation may be relevant to their situation (Zwaan, van Voorst, & Mastenbroek, 2016). Furthermore, the comparability of evaluation findings (Schoenefeld et al., 2018) becomes a core issue when the goal is to carry lessons from one governance centre to another. Related to the issue of comparability, Feldman and Wilt (1996) have argued that

> To ensure that states and other regional jurisdictions can be equivalently evaluated on their progress in achieving these [climate] goals, some means must be developed to collect valid energy and emissions data across jurisdictions and—equally important—to ensure that these data measure the same things in the same way (p. 49).

Thus, the extent to which an evaluation includes metrics that allow comparison across governance centres matters in this respect.

And yet reverting back to the debate on idiosyncratic evaluation criteria and generalization (see above) raises key and difficult questions about comparability and thus learning opportunities (Schoenefeld & Jordan, 2017; Schoenefeld et al., 2018). A combination of providing both contextual analysis that takes into account contextual effects, but also some more general criteria or metrics that enable comparison seems of order. Aligica and Sabetti (2014b) draw on Elinor Ostrom to explain that this may be done by conceptualizing and researching 'basic units' of policy or interaction that appear across multiple contexts, without aiming to make broad and sweeping generalizations that are unlikely to hold. True to the argument that supposed panaceas are unlikely to work (see E. Ostrom et al., 2007), the polycentric approach would highlight the importance of context in determining to what extent lessons can 'travel.' In line with the discussion on context above, in order to be a useful tool in fostering interactions between governance centres, climate policy evaluations would have to carry some level of contextual information in order to enable lesson drawing in context. The idea that evaluation can generate knowledge that travels between different governance centres is relatively new and has surprisingly been little discussed in evaluation literatures.

Then there is the potential interaction between formal and informal evaluation activities. As discussed earlier, scholars have developed the distinction between formal and informal actors in evaluation. But how do the 'formal' and the 'informal' spheres of policy evaluation interact, if at all? There has been a growing interest in informal governance (Helmke & Levitsky, 2004) with a particular focus on the EU in recent years (Christiansen & Neuhold, 2013; Kleine, 2013). These literatures suggest that the interaction between formal and informal institutions may be "complementary, accommodating, competing [or] substitutive" (Helmke & Levitsky, 2004, p. 725). In the complementary case, informal institutions may fill gaps left by formal institutions, whereas in the accommodating variant, informal institutions may influence the way formal institutions work without seeking to do away with them. By contrast, in

competing or substitutive cases, informal institutions ultimately seek to replace formal institutions (Helmke & Levitsky, 2004). However, particularly when formulating policy recommendations, 'informal' does not necessarily mean 'disorganized' or 'worse' (Guha-Khasnobis et al., 2006). In sum, theory suggests that there are numerous ways in which informal and formal institutions may interact. In studying evaluation in polycentric systems, this distinction is crucial, because it begins to identify the multiple actors that could be involved in evaluation, and goes beyond assuming that the main site of evaluation is necessarily government.

Evidence suggests that actors do pay attention to one another on climate policy questions. For example, *The Economist* wrote in November (2014) that

> Officials in California, for example, made several fact-finding visits to Brussels to investigate the EU's emissions-trading regime when preparing their own […]. Before its launch two years ago the Californians told sceptics that they had learned important lessons from the European example—even if these were largely about what to avoid.

Earlier on, the EU had looked to the USA for key lessons from sulphur dioxide trading for their own emerging carbon dioxide emissions trading scheme; an example of this activity is a 1999 report by the EEA, which looks at several procedural issues and the overall US experience with emissions trading systems (Mangis, 1998). Such effects have been studied much more systematically in relevant policy diffusion literatures. In their review of these literatures, Jordan and Huitema (2014) explain that states may have significant incentives to interact, with a desire to learn as one of the headline motives.

But in addition to these points of learning, policy evaluation may also aid governance centres to hold each other to account (as is the hope of the transparency mechanisms in the Paris Agreement), and potentially also allow actors within governance centres to contribute to accountability mechanisms. In addition, knowledge flowing from evaluation may, to a certain extent, also enable competition between governance centres (see V. Ostrom et al., 1961) by for example providing a basis for benchmarking. However, the extent to which this happens with a view to accountability

and competition remains an open question, as the political and potentially strategic nature of policy evaluation (see above) may also make evaluation actors reluctant to publicize their findings, particularly when they describe key factors that drive success.

Linked to the above discussion is the question of *intended* evaluation use. While knowledge use in public policy is a widely debated topic in political science (e.g., Albaek, 1995; P. Haas, 2004; Radaelli, 1995; Rich, 1997) for space and practical reasons this thesis considers the more circumscribed *intended* target audiences (and thus potential users) of an evaluation. Intended evaluation users are often policy-makers, although some evaluations may be conducted for accountability or even strategic purposes. Prominent evaluation approaches focus in particular on utilization. For example, Patton (2008, p. 37) takes the view that "the focus in utilization-focused evaluation is on *intended use by intended users"* (emphasis in original). This statement thus begs the question who the intended users are, but to date, there is virtually no empirical evidence to address this question, especially for climate change policy. In these conceptualizations the users of evaluations tend to come from fairly small circles. By contrast, the polycentric approach would envision uses of evaluation that go well beyond a relatively narrow set of users, such as the creators of a policy, or those who are being affected by it.

Currently, evaluation is typically done by policy-makers themselves (either in-house or commissioned) or by those who have a stake or interest in the outcomes of a particular policy. In polycentric systems, one key difference that has so far received little attention is that the circle of potential evaluation users widens to include others in governance centres that do not have a direct stake in the outcome of a particular policy, but who may be able to benefit from insights generated by an evaluation elsewhere (related to learning, see above). Another function is to provide some accountability in governance settings where traditional accountability chains have been weakened or no longer exist (see Bäckstrand, Zelli, & Schleifer, 2018). In this understanding, evaluation becomes in effect a public good, which is non-exclusionary (if evaluations are public) and non-rivalrous (the use of insights by one user does not preclude another one from benefitting from the insights). In this regard, policy evaluation in polycentric

governance systems potentially departs from current understandings of policy evaluation as the scope of possible evaluation users expands.

2.6 Conclusion

The previous sections have endeavoured to make a theoretical case for examining the importance and actual roles of policy evaluation in facilitating climate governance by contributing to the shift from polycentricity to polycentrism. They show that literatures on polycentric governance and policy evaluation have already engaged with concepts that are highly relevant, yet often ill developed and with virtually no connection to the body of literature on the other side. The respective debates have by and large taken place in relatively self-contained, and often self-referential, scholarly communities with their own set of dedicated journals, conferences, and networks. For example, evaluation literatures have already debated the role of context in evaluation, as well as the role of multiple actors and—to a much lesser extent—the notion of interacting governance centres. But to date there is a severe paucity of studies that consider all these factors simultaneously. Insights from this kind of integrative research across different factors could help shed light on the potential and actual roles of (climate) policy evaluation in polycentric governance systems. The above review shows how information provision via policy evaluation is in many ways implicit in Ostrom's polycentric governance theory, but its precise role and to what extent this happens in practice have yet to be explored.

This chapter set out to identify the basic theoretical building blocks of polycentrism, which as a theory contains both normative and positive elements. It shows that these foundational insights are that context matters in governance, that actors can and sometimes do self-organize to muster governance solutions and that interaction between otherwise independent governance centres appears indispensable in order to move from polycentricity to polycentrism (see Chapter 1). Bearing in mind arguments about scale in governance, the chapter shows that we can draw key theoretical insights from monitoring studies in common pool resource governance systems in order to

conceptualize the role of policy evaluation in polycentric governance systems. Crucially, policy evaluation can potentially make significant contributions to the emergence of polycentrism, but in order to do so, it must exhibit certain features outlined in the sections above. Moving forward, this newly developed theoretical approach thus provides some yardsticks against which we can evaluate the practice of climate policy evaluation in the next chapters.

Chapter 3 Climate Policy Evaluation: the EU Level, Germany, and the UK

3.1 Introduction

This chapter reviews existing literature on the historical evolution and current functioning of (climate) policy evaluation activities at the EU level, in Germany, and in the UK. In each case, this chapter reviews the emergence and drivers of policy evaluation in general and then provides an overview of the institutions and organizations that support evaluation, including key evaluation actors. This is followed by an overview of the rise and nature of environment and climate policy evaluation at the EU level, as well as in Germany, and in the UK. Finally, this chapter assesses the current state of knowledge against the foundational ideas of polycentric governance, namely self-organization and context in evaluation, as well as interaction between governance centres through evaluation (see Chapter 2). In sum, the chapter assesses the current state of the literature, and it exposes key gaps in our knowledge and understanding.

3.2 Evaluation at the EU level

3.2.1 Historical development

Following growing pressures on governments since the 1960s to demonstrate the effectiveness of their various policies, the EU very much stepped up its activities to evaluate the outcomes of structural and cohesion funding in the 1980s and 1990s (Stame, 2003; Summa & Toulemonde, 2002). At the same time, policy evaluation received a boost from 'New Public Management' (NPM) thinking and reforms with their focus on value for public money and accountability in the 1990s (Stame, 2003).

This manifested in several EU member states and finally at the EU level in the 'better regulation' (Radaelli, 2007) and, since 2010, in the 'smart regulation' agenda (European Commission, 2010). More recently, other factors such as the financial crisis, the recession and Euroscepticism have added additional pressures to evaluate policies (Stephenson, 2015). In sum, evaluation has been discussed and addressed at the EU level for various decades.

3.2.2 Actors and institutions

There are a range of formal, state-like actors at the EU level that in one way or another participate in and contribute to policy evaluation. Key actors include the European Commission (EC), the European Environment Agency (EEA), and, to a lesser extent, the European Parliament, the European Council, and the European Court of Auditors. This section reviews existing knowledge and literature on the role of these institutions in policy evaluation.

European Commission. The historical pressures toward evaluation explained in the previous section generated significant—and growing—evaluation demand and corresponding activity levels in the European Commission, which is located at the centre of EU policy-making as the guardian of the EU treaties, as well as the sole policy initiator (Peterson & Shackleton, 2012). Back in 1996, Nordic Commissioners pushed for a communication in order to spread evaluation across all Commission Directorate-Generals (DGs) (European Commission, 1996; see also Summa & Toulemonde, 2002). This initiative generated a decentralized evaluation 'system' within the Commission where each DG coordinates its own evaluations (Hojlund, 2015). This structure continues to be in place at the time of writing (European Commission, 2013), and it has led to varying levels of evaluation capacity in individual DGs (van Voorst, 2017). Given limited staff capacity, the Commission in practice typically outsources policy evaluation (European Commission, 2013; Summa & Toulemonde, 2002). Even though internal evaluation standards have provided a broad framework since 2002, in practice the

standards became a rather loose set of 'guiding principles' (European Commission, 2002). This leaves considerable discretion to those who contract evaluations. As Stame (2008, p. 124) writes, "at the EC level, evaluations are conducted according to predetermined and once commissioned generally inflexible terms of reference, established by the commissioning DG." For example, in the field of cohesion policy evaluation, this means that when evaluations are commissioned, the choice of methodology is largely up to the individual contractor (Batterbury, 2006). Furthermore, the Commission standards are largely process-oriented (including the establishment of a steering group for each evaluation) so that

> [a]s to the content and type of evaluations, the rule is not one of standardization. On the contrary, the Commission emphasizes that evaluation projects should be tailored to the objectives and delivery mechanisms of the policy or program concerned. (Summa & Toulemonde, 2002, p.415)

Taken together, the European Commission has been organizing its internal evaluation activities decentrally since the 1990s. In order to provide more cumulative insights, DG Budget produced ten 'Annual Evaluation Reviews' between 2000 and 2009.[6] However, even in 2013, Per Mickwitz asserted that policy evaluation in the EU "[…] is still not well institutionalized […]" (Mickwitz, 2013; see also Mickwitz, 2003).

Evaluation literatures recognize that contracting evaluations generates a range of potential issues regarding the (perceived) independence of evaluation results (see Chapter 2). The European Commission has not escaped these dynamics. In some instances, it has been criticized heavily for producing internal policy evaluations that selectively present information to suit its own political interests (Versluis et al., 2011, p. 224). Furthermore, in some areas the Commission frequently relies on the Member States to conduct evaluations decentrally. According to Batterbury (2006, p. 184),

[6] http://ec.europa.eu/smart-regulation/evaluation/documents_en.htm

> In theory, the decentralization of evaluation should give it a greater proximity to the stakeholders and offer greater opportunities for locally sensitive evaluation design. In practice, evaluation capacity is highly differentiated across the EU territory, reflecting differing evaluation traditions, experience and resources […].

The European evaluation landscape that developed from these early beginnings became highly variegated, with different 'evaluation cultures' in different countries (Furubo et al., 2002; Jacob et al., 2015; Polverari & Bachtler, 2004). In general, Versluis et al. (2011, p. 224) write that

> […] EU evaluation culture is political and pluralistic, characterized by a variety of organizations willing to pay significant sums of money to finance research that may produce data in support of their political views.

This has generated an overall European evaluation landscape that is pluralistic, a feature that evaluation scholars describe as a sign of more advanced evaluation activities (Jacob et al., 2015), but which is certainly a challenge for the Commission in seeking to bring together insights from across Europe.

European Environment Agency. The second key actor with an increasingly important role in policy evaluation is the European Environment Agency (EEA). Overall, scholars agree that from the start in the 1990s, the evaluation role of the EEA and especially its relationship to the European Commission (see above) has been contested (Martens, 2010; Zito, 2009). Early on, EU Member States and Members of the European Parliament disagreed over the EEA's role, with some of the former (in particular the UK and Spain) advocating a 'data collection' or monitoring role for the EEA, whereas some members of the European Parliament preferred inspection powers in order to hold national policy implementers to account (Zito, 2009) – so a more explicit evaluation role. For example, an Institute of European Environmental Policy (IEEP) evaluation in 2003 reported on letters from two Directors-General of DG Environment, which indicated that

> The development of policies, implementation reviews, policy evaluations and recommendations were the responsibility of the Commission alone. The Agency should not get 'sidetracked by the more glamorous but rather sensitive hot political issues.' (IEEP & EIPA, 2003, p. 39)

Although a compromise ultimately weaved these divergent views into the regulation that underwrites the EEA (Martens, 2010), there was initially a focus on monitoring and data collection. However, over time, the EEA became more involved in developing *ex-post* evaluation methodologies without conducting the evaluations themselves (see IEEP & EIPA, 2003). This shift may signal a gradual acceptance that the EEA's role is evolving in a more policy-analytical direction. While in the early days, some actors such as DG Environment in the Commission felt threatened by the EEA, these tensions eased over time, with the EEA building much closer relationships with the Commission, the European Parliament and the Council (Zito, 2009).

The European Parliament. In contrast to the European Commission, the European Parliament has generally engaged less in policy evaluation (Stern, 2009; Hojlund, 2015). This is especially true for environmental policy (Mickwitz, 2013). Historically, it has lacked internal evaluation capacity. However, in a 2008 speech, the then Chair of the Committee on Budgetary Control in the European Parliament argued in a debate that

> A parliamentary evaluation function would try to make transparent to the citizens what they are getting for their money. This objective is quite different from predominantly helping the policy system to justify how it is spending the money. Evaluation should not be brought down to a 'management tool' as is now the case in the EU system. (cited in Stern, 2009, p. 70)

New evaluation capacity has subsequently slowly materialized within the European Parliament. In 2012, it created a 'Directorate G for Impact Assessment and European Added Value,' which comprises various units that address *ex-ante* impact assessments

and *ex-post* evaluations (Poptcheva, 2013). Academic researchers have recognized this development, confirming that "[…] the Parliament… [has] sought to strengthen… organizational capacity to perform better, by increasing resources for research and placing greater emphasis on results and impact assessment" (Stephenson, 2015, p. 83).

European Council and Court of Auditors. Similar to the European Parliament, the European Council has been a weak evaluation actor (Stern, 2009). As Mickwitz (2013) details, it has had no role in environmental policy evaluations. By contrast, the European Court of Auditors appears to have a growing role in policy evaluation, especially as it extends its remit from financial auditing to evaluating broader policy effects. Some of these evaluations concerned environmental policy (Mickwitz, 2013). In a similar vein, Stephenson (2015) describes a development towards performance evaluation in the form of 'special reports.' To do so, the court has introduced a new diploma course, given that performance evaluation requires a much broader skillset than relatively straightforward and streamlined financial audits (Stephenson, 2015). However, the special reports also differ from other evaluation reports, because the court often evaluates policy-making across many sectors and because it has access to privileged financial information that may not be available to other evaluators (Stephenson, 2015). In sum, the European Court of Auditors is becoming more important in evaluation.

Similar to Germany and the UK, much less is known about evaluation outside the formal EU institutions or by extension of argument, beyond the state. European evaluators founded the European Evaluation Society (EES) in 1994 (Bemelmans-Videc, 1995), which brings together national evaluation organizations and serves as an international platform for information exchange and learning through conferences, newsletters and journals (Schröter, 2007). However, the association has not been very visible on environment and climate policy evaluation and relevant literatures lack debates on its interactions with the governmental actors.

3.2.3 Climate policy evaluation

Policy evaluation activities in the environment sector – the traditional 'home' of climate policy in the EU - have been slower to develop than in other sectors, even in general evaluation forerunner jurisdictions such as the US (Knaap & Kim, 1998). However, by 2001, the European Environment Agency reported significant, if still to-be-improved, evaluation activities in the environmental sector around the EU (European Environment Agency, 2001). In 2002, the EU thus called for improved *ex-post* environmental policy evaluation in its sixth ten-year Environmental Action Programme (Mickwitz, 2003).

The Rio Earth Summit in 1992 became a watershed moment for climate policy because it laid some of the foundations for later climate policy evaluation in the EU and beyond. After the issue of climate change had risen to the attention of international policy makers in the late 1980s (T. Rayner & Jordan, 2013), the EU found itself under pressure to attend this summit with concrete policy proposals. Having failed to agree on an EU-wide carbon tax, European negotiators instead focused on creating systems for greenhouse gas monitoring at the international level, which were agreed at the summit (Bodansky, 1993; Jordan et al., 2010; Yamin & Depledge, 2004, p. 327). This, in turn, led the EU to adopt a Monitoring Mechanism for greenhouse gases and later policies and measures in 1993 (Haigh, 1996; Hyvarinen, 1999; Schoenefeld et al., 2018). Successive revisions of that mechanism—usually in the context of international negotiations—attempted to refine the mechanism in order to collect more data on individual policies in addition to the national-level greenhouse gas inventories (Farmer, 2012; Hilden et al., 2014). These developments laid the groundwork for some more basic data collection through a formal monitoring system, which would later allow evaluators to generate more wide-ranging evaluations.

Aside from the discussion on the Monitoring Mechanism, there is little systematic knowledge on the engagement of the various EU institutions in the area of climate policy evaluation. While an early meta-analysis found that evaluators at the EU level are the most productive across the EU (Haug et al., 2010), most scholarship has focused on evaluation institutions in general and on evaluation in other policy fields, generating an

immense knowledge gap on the nature of climate policy evaluation. This remains true for the case of the European Commission, although scholars have detected a tendency by the Commission to work towards harmonizing climate policy evaluation by encouraging independent reviews of evaluation methodologies and practice around the member states (AEA, ECOFYS, Fraunhofer, & ICCS, 2009; Hildén et al., 2014; Öko-Institut et al., 2012; Schoenefeld et al., 2018). This paucity of knowledge contrasts with an overall assessment of environment and climate policy evaluation in the EU by the EEA, which argues that "[t]he evaluation of environment and climate policies is, today, a well-established discipline" (European Environment Agency, 2016, p. 4).

An exception of sorts may be the EEA itself. Climate policy—and at times its evaluation—is perceived as a central work area for the EEA. For example, a 2008 evaluation of the EEA found that

> The majority of the Commission officials interviewed consider climate change has been a core element of the EEA's work. […] The EEA's inventory on GHG emissions as well as its monitoring of progress in GHG emissions and projections in the EU via the European Topic Centre for Air and Climate Change are valuable for DG Environment. (Technopolis, 2008, p. 35)

Furthermore, a 2013 report shows that people believed that climate change is one of the policy areas where the EEA has most impact (COWI, 2013, p. 47). In a similar vein, the 2008 evaluators found the EEA receives comparatively high press coverage on climate change and that it is one of the top interests of its website users (Technopolis, 2008). Paradoxically, while the Commission appears to value perceived EEA independence in climate change reporting, because it fears legal challenges through the UNFCCC (COWI, 2013, p. 35), it simultaneously continues to reject a more active role for the EEA in policy analysis, even though this is an area where the EEA continues to become more engaged. For instance, the Technopolis (2008) evaluation identified a clear appetite by the EEA's Management Board, as well as National Focal Points, for a stronger policy measure effectiveness evaluation role on climate change (p. 41). This may be especially because these two groups of people consider climate policy measure

effectiveness evaluation one of the EEA's established roles (p. 40). Member states may indeed have more data available that they are not willing to supply to the EEA via the Monitoring Mechanism. As the COWI (2013) report explains,

> While the Member Countries appreciate the need to establish pan-European datasets and assessments, the value of these is considered modest in the national context. While it is recognised that it is valuable for benchmarking and learning, most Member Countries still emphasise that they have more detailed data at national level and thus derive national assessment on a different basis. (p. 23-24)

However, the Commission position remains largely unchanged: as the Technopolis (2008) report explains, "[t]he overall view [by the Commission] was that the Agency might have a role in developing tools and methodologies in some cases but not in actual policy assessment […] (p. 42, see also IEEP & EIPA, 2003, p. 49). Thus, the role of the EEA in climate policy evaluation may be described as ambivalent, and evolving. While it is clear that it is heavily bound up in the wider dynamics of European politics, there is today no systematic overview of the climate policy evaluation outputs by the EEA.

Even less is known about the level and nature of climate policy evaluation in the European Parliament and the European Court of Auditors where an analysis of their outputs, particularly with regard to climate change policy, does not yet exist. Little is known about the outputs of these formal evaluation actors in the EU. However, in a recent study, Mastenbroek et al. (2015) highlights that "[*ex-post* legislative evaluation] is primarily a matter of legislative obligation instead of own initiative" (p. abstract).

3.3 Evaluation in Germany

3.3.1 Historical development

While policy evaluation became increasingly prominent in the late 1960s in Germany (Derlien, 2002), sometimes described as an 'explosion' in the activity (Lange, 1983), the public budget crises of the 1970s considerably reduced the enthusiasm for evaluation

(Stockmann, 2006). Since the 1990s, evaluation has gained importance again, driven by factors such as greater legitimacy demands towards government (Brandt, 2009), increased use of EU structural funds following Germany's reunification in 1990 (Taylor, Bachtler, & Polverari, 2001) and the need to restructure and in some cases shut down decrepit institutions in East Germany (Derlien, 2002). Furthermore, the philosophy of New Public Management (NPM) with its concern for efficiency and effectiveness in governmental affairs arrived in Germany in the 1990s and became another driver of evaluation (Derlien, 2002; Löwenbein, 2008; Pattyn, Van Voorst, Mastenbroek, & Dunlop, 2018). Overall, while frontrunner jurisdictions such as United States have certainly influenced evaluation in Germany (Struhkamp, 2007), some have argued that it never reached the level found in the Anglo-American sphere (Stockmann, 2006). A more recent study however points towards relatively mature evaluation activities within Germany (Jacob et al., 2015).

3.3.2 Actors and institutions

The general level of institutionalization of evaluation in governmental institutions remains relatively low in Germany (Jacob et al., 2015). While multiple actors at various governance levels became involved in evaluation (or 'success control,'[7] as it was sometimes called in German), few established in-house evaluation capacities (Struhkamp, 2007). The federal government has emphasized the need for evaluation rhetorically, but this has not yet led to a firm establishment and use of evaluation in government (Stockmann, 2006), where evaluation remained highly fragmented even within individual ministries (Derlien, 1990; Löwenbein, 2008). There is no central institution that manages evaluations in Germany (Duscha et al., 2009). An exception may be the creation of an 'Independent Evaluation Agency' by the German Parliament (Jacob et al., 2015).

[7] Erfolgskontrolle

The German Federal Court of Auditors[8] has also published two studies in 1989 and 1998 that highlighted the need for evaluation, and to this day courts are among the most ardent advocates of evaluation in Germany – especially in the area of evaluating the use of public funds (Stockmann & Meyer, 2014, p. 48; Struhkamp, 2007). In sum, while the nature of evaluation activities varies by sector (Beywl, Fabian, & Widmer, 2009; Derlien, 2002), they are highly decentralized in Germany, especially in the case of environment and climate policy evaluation (Duscha et al., 2009). A general lack of coordination of evaluation activities has fostered many sectoral, but few comprehensive, evaluation activities (Stockmann, 2006).

This is also true for evaluation developing outside the echelons of government or other state institutions. The field of professional evaluators was equally slow to develop – but the founding of the German Society for Evaluations (DGeEval)[9] in 1997 appeared to bring together many decentralized activities, as the association boasted almost 800 individual and institutional members by 2014. In a more recent study, Brandt (2009) documents a professionalization process in German evaluation activities, but also limits thereof—expressed for example by a reluctance of members of the German Evaluation Society to see evaluation as a separate profession. The founding of the German 'Zeitschrift für Evaluation'[10] in 2002 generated a new platform for sharing evaluation experience and knowledge. Importantly, the DGeEval publishes evaluation standards – which appeared in 2008 in their fourth incarnation.[11]

Some knowledge has also emerged on the nature of policy evaluations from Germany. Löwenbein (2008) estimated that total spending on evaluation in Germany amounted to 134 Million Euros – far from insignificant and considerably more than the 45 Million Euros that the European Commission spends per annum on evaluation (Hojlund, 2015). Furthermore, Löwenbein (2008) estimated costs of about 100.000

[8] 'Bundesrechnungshof' in German.

[9] http://www.degeval.de/nc/home/

[10] Journal of Evaluation (http://www.zfev.de/)

[11] http://www.degeval.de/degeval-standards/

Euros per evaluation in Germany. Finally, existing literature suggests that Germany has a long tradition of commissioning evaluations—in part because of the decentralization discussed above—(Struhkamp, 2007), with private consultancies as the prime evaluation producers (Löwenbein, 2008) but also a vibrant and on-going debate on what one means by 'inside' and 'outside,' as well as 'self-evaluation' (Struhkamp, 2007) – a point to which this thesis returns in later chapters.

3.3.3 Climate policy evaluation

By and large, environment and climate policy evaluation practices emerged in Germany in the 2000s (Duscha et al., 2009). However because many of the institutes that evaluate environmental policy have their roots in the natural sciences, their work has to date only benefitted from very limited engagement with insights on evaluation from the social and policy sciences (Duscha et al., 2009; Kaufmann & Wangler, 2014). For example, environment and climate policy evaluation actors generally do not use the evaluation standards by the DGeEval, and they typically do not attend the conferences of the association either (Duscha et al., 2009). Other authors, such as Wörlen, Rieseberg, & Lorenz (2014, p. 2) came to similar conclusions, stating that "in the energy and environment field, the evaluation tradition in Germany is so far rather weak." However, this is different in the case of the Energiewende and other climate-related policies, which are increasingly dynamic sites of evaluation, including by governmental and non-governmental actors (Kaufmann & Wangler, 2014; Wörlen et al., 2014). The German 'Integrated Energy and Climate Program'[12] by the German federal government, which has been in place since 2010, prescribes a regular monitoring and evaluation exercise, starting in 2010 (Doll et al., 2012). The federal government commissioned various research institutes and individuals to assist in designing this monitoring program (Doll et al., 2012). However, as Wörlen et al. (2014, p. 2) point out, this monitoring process could still be improved, for example by providing better and more streamlined

[12] http://www.bmwi.de/DE/Service/gesetze,did=254040.html

indicators, and by defining the target audience more clearly and by better communicating findings with the public. Importantly, in response to the perceived shortcomings of the governmental reporting, various other and non-governmental actors have prepared their own indexes and evaluation approaches, although some of them turned out to be rather short-lived (Wörlen et al., 2014).

Beyond the discussion of specific monitoring processes, Germany has also been identified as one of the most productive origins of climate policy evaluations in the EU – roughly one fifth of all evaluations identified by Haug et al. (2010) came from Germany. However, which kinds of actors funded and/or conducted these climate policy evaluations is not reported in relevant papers on the ADAM project (see also Huitema et al., 2011) and there is very little information on the content of the evaluations, and their methodologies. Such gaps in the knowledge are in line with Duscha et al. (2009), who recommend collecting and carefully analysing evaluations in the climate and energy sector and beyond.

3.4 Evaluation in the UK

3.4.1 Historical development

The United Kingdom has been an early adopter of policy evaluation, which first emerged from concerns over policy effectiveness in the context of new public management reforms dating back to the 1960s (Gray & Jenkins, 2002). In contrast to Germany, the United Kingdom is a highly centralized state, and this has affected policy evaluation endeavours. By and large, central government and its constituent parts drove the first attempts to evaluate. For example, the Treasury established the 'Public Expenditure Survey' in the 1960s or the 'Program Analyses Review' in the 1970s, but these initiatives never fully materialized given internal resistance (Levine, 1984). Later, evaluation was mainly seen as a tool to avoid overspending and allocate scarce resources, especially after an International Monetary Fund (IMF) bailout in the mid-1970s (Pattyn et al., 2018), although during the Thatcher years there was an attempt to

decentralize evaluation into the departments and issue evaluation guidelines (Jenkins & Gray, 1990).

Since the late 1990s, much of the academic discussion on evaluation-related activities in the UK sailed under the banner of 'evidence-based policy-making,' as this was one of the major tenets of the Labour governments between 1997 and 2007. Tony Blair's 'New Labour' government saw evaluation as a key component of a 'third way' of pragmatic, evidence-based policy-making beyond political ideology and subsequently issued a range of evaluation guidance documents (Pattyn et al., 2018; Sullivan, 2011). Wells (2007, p. 27) argues that, as a function of evidence-based policy-making under New Labour,

> evaluation has become a more widely accepted part of the policy making process, more frequently and knowledgably used by central government and local and regional agencies. For instance, evaluation designs have become more sophisticated with greater use of a range of data, including longitudinal elements as well as theoretically based approaches.

While much more evaluation funding became available during this period and while evaluators became a more important source of policy advice, several characteristics limited evaluation. Even though there was an emphasis on 'theory-based evaluation,'[13] scholars admonish that evaluation practitioners focused mainly on rationalistic approaches to evaluation, and tended to avoid more discursive ones (Sullivan, 2011). This, in turn, has partially stifled the potential role of formal evaluation in the United Kingdom. However, the theory-based approach also led to a greater attention to context, because this method necessitates theorizing individual programs, rather than the policy system at large (Sullivan, 2011). To what extent such developments appear in climate policy in particular remains very much an open question.

[13] The idea that evaluation should be based on programme theory, that is, programme-specific theories of cause and effect.

This brief history shows that the waxing and waning of policy evaluation in the UK has always to a certain extent depended on the powers that be at any one point in time. Conservative governments often believed that market forces were enough as an evaluation force, and as a result "[i]t is not hard to argue that systematic policy and program evaluation in the U.K. became about as impoverished in the late 1980s and early 1990s as at any time since 1945" (Gray & Jenkins, 2002, p. 135). Tony Blair's Labour government proved more receptive to evaluation, but it had clear preferences for certain types of evaluation.

3.4.2 Actors and institutions

The UK has experimented with various evaluation institutions since the 1960s (Gray & Jenkins, 2002; Jenkins & Gray, 1990). In contrast to Germany, there was initially a much stronger drive to centralize evaluation in national-level institutions and to provide national-level evaluation guidelines as evaluation dispersed into the departments (Gray & Jenkins, 2002, p. 135; Jenkins & Gray, 1990). Over the last two decades or so, there has subsequently been a general trend towards standardizing and prescribing evaluation methods, which was set out in general terms in the Green Book on Appraisal and Evaluation in Central Government (HM Treasury, 2003) and, as an extension thereof, the Magenta Book on Guidance for Evaluation (HM Treasury, 2011). These documents aim to streamline and define evaluation not only in government, but also beyond as the authors emphasize in the introduction to the Magenta Book (HM Treasury, 2011). Generally, they emphasize economic aspects of evaluation.

Often changes in government meant significant institutional swings that have stifled attempts to build enduring evaluation institutions in order to generate long-term insights (Gray & Jenkins, 2002). Pollitt (1993) noted early on that

> Perusal of the history of the last thirty years reveals that policy evaluation [in the UK] has never found a secure or permanent home near the heart of a (relatively centralised) state. (p. 354)

For example, the UK never created the types of more long-standing policy research organizations like those found in Germany or in the USA, and what existed by way of evaluation institutions tended to be short-lived, financially instable, and highly dependent on the (political) support of the powers that be at any one point in time (Parsons, 2007; Pollitt, 1993). Therefore,

> […] the activity of policy evaluation has been a precarious one, and its practitioners have had to become hardened to their work frequently being ignored or ridiculed by politicians and the mass media. Britain still lacks a political culture which is broadly supportive of deep analysis and assessment of complex policy problems (Pollitt, 1993, p. 354).

The New Labour government under Tony Blair created a range of units and institutions aimed to further rationalistic, evidence-based policy-making (Parsons, 2007).

While evaluations in the early decades of the practice in the UK were often conducted internally (Levine, 1984), this practice changed in later years. A recent UK-based survey suggests that when the government funds evaluations, a number of (academic) evaluators felt that government officials sought to influence evaluation results (Hayward et al., 2013). This may affect the independence of evaluation results, and could make it difficult to reflect critically on the program goals (Huitema et al., 2011), or assess co-benefits that may policies may generate (E. Ostrom, 2010c). But more recent assessments have emphasized that some evaluation institutions appear to have crystallized in the intervening decades. Today, various state institutions play relevant roles in the UK. Aside from central government with various ministries in leading roles, select committees in Parliament, as well as the National Audit Office (NAO) have been cited as strong producers of evaluations (Pattyn et al., 2018). In sum, Pattyn et al. (2018, p. 7) argue that in the late 2010s, "[t]he UK's institutional arrangements for conducting and disseminating evaluations are strong."

In addition to these mainly state institutions, professional evaluators have also formed the UK Evaluation Society in 1994 (Risley, 2007). The association has between 200 and 300 members (UK Evaluation Society, 2013), it holds regular conferences,

publishes several journals and has produced evaluation guidelines, which it publishes on its website. However, while Widmer (2004) notes that UK evaluators were relatively late in discussing evaluation standards, he also acknowledges that the literature has not discussed interactions between the UK Evaluation society and the governmental efforts to evaluate, or more specific aspects of environment and climate policy evaluation.

3.4.3 Climate policy evaluation

The UK is one of the most active and advanced evaluators of climate policy (AEA et al., 2009; Haug et al., 2010; Huitema et al., 2011; Öko-Institut et al., 2012). A major meta-analysis identified the UK as one of the most prolific producers of climate policy evaluations (Haug et al., 2010). In line with the general attempts to streamline evaluation in the UK (see above), in 2010 the UK government published specific guidance on how to estimate the effect of policy on greenhouse gas emissions (HM Treasury & DECC, 2010). However, in the intervening decade, little has been written on the evaluation of climate policy in the UK, constituting a key research gap.

3.5 Self-organization, context, and interaction

This section brings together insights from the above review and extends them with a view to the three foundational ideas of polycentric governance that Chapter 2 specified.

3.5.1 Self-organization

The literature review above suggests that while existing evaluation literatures focus almost exclusively on formal (i.e. state-led) evaluation institutions, actors, and outputs (for a recent example, see Mastenbroek et al., 2015), there is a recognition among scholars in the EU that "[v]arious multi-level stakeholders will conduct their own

evaluations and choose from a mix of qualitative and quantitative data" (Stephenson, 2015, p. 82). However, very little else is known about informal evaluation activities outside the more familiar (state) institutions. This underexplored area thus constitutes a vast research gap. As Chapter 4 explains in greater detail, self-organization may engender a range of other, non-state actors in climate policy evaluations. And yet, extant literature on policy evaluation in Germany, in the UK and at the EU level has almost nothing to say about the role of self-organization in climate policy evaluation, let alone any interactions between state and more self-organized policy evaluation activities. This lack of knowledge is puzzling especially because the bulk of the climate policy evaluations analysed in a recent meta-analysis was done by informal actors (Huitema et al., 2011). In other words, a significant number of actors appear to conduct evaluation outside formal state structures. However, given that academic papers were included as 'evaluations' and given that universities and research institutes were one of the most productive locus of evaluation by author affiliation, it remains an open question to what extent we can see self-organizing tendencies in evaluation, meaning that those directly or indirectly affected by a policy also conduct their own evaluations (Huitema et al., 2011).

What we do know from this analysis, however, is that 'informal' or supposedly more independent evaluations exhibited a somewhat greater, but not overwhelming, propensity to critically question the goals of the original policy they evaluate, that is, reflexivity (Huitema et al., 2011). This may, in turn, also come with greater attention to the specific context of a climate policy under evaluation, but the current literature provides no insights on the extent to which this may be happening.

3.5.2 Context

This literature review could not unearth any studies that may have assessed attention to contextual factors in climate policy evaluations in the EU directly in the sense that evaluation scholars have discussed it (see Chapter 2). However, there are several elements in the literature that point in useful directions. One issue is the extent to

which evaluation actors in the EU have attempted to harmonize evaluation practice. Echoing Ostrom, context-sensitive policy evaluation is hard, or perhaps even impossible, to standardize. But this is not to say that it hasn't been tried in other policy fields. Writing on the evaluation of structural funds, Batterbury (2006, p. 187) suggests that

> The decentralization approach should ensure that greater contextual relevance is integrated into the evaluation process. However, the heavy emphasis on both quantifying impact and measuring performance misses the key question: why things work (or not) in specific contexts.

In other words, at least with regard to cohesion policy, one of the early fields of policy evaluation, current methodological aims to standardize evaluation practice appears to stand in the way of paying closer attention to the context in which a policy may operate. This is especially relevant for climate policy making because, as Elinor Ostrom suggests, co-benefits of climate policy may contribute significantly to success in polycentric settings (E. Ostrom, 2010c). Simple indicator-based monitoring and evaluation—such as that practiced by the EEA in the context of the Monitoring Mechanism (see above)—may be unlikely to capture unexpected co-benefits or provide data on the full range of policy impacts, given that the latter likely vary with different policies and in different contexts. This chapter has disentangled various standardization attempts in (climate) policy evaluation, for example the methodological studies of the European Commission, or the evaluation guidelines issued by the UK government. But by the same token, a significant part of policy evaluation appears to be fashioned in a decentralized manner (especially in the European Commission), which may foster more attention to context. But the extent to which this is so has not been sufficiently studied.

Second, as Chapter 2 explained, scholars have argued that using multiple methods may help in capturing multiple contextual policy effects. In the context of European spatial planning policy, Dabinett and Richardson (1999) explain how the 'hegemony' of economics and a 'pro-integration bias' tends to make evaluation a tool of powerful interests in policy-making, rather than a means to engage in pluralist democratic discussions and debates about policy alternatives. A recent meta-analysis suggests that

more than half of the 259 climate policy evaluations at the EU level and from various member states used just one evaluation methodology (Huitema et al., 2011). This lack of methodological diversity again calls into question whether current climate policy evaluation practices capture the full range of co-benefits and thus contextual effects of various policies. Climate policy evaluation literature entirely miss out on gauging to what extent evaluations pay attention to the various contextual dimensions that Chapter 2 describes, such as political, spatial, or scientific aspects. This constitutes a knowledge gap in extant literatures.

3.5.3 Interaction

Last, to what extent climate policy evaluations engage with and draw on experiences and activities in other governance centres has not been analysed in this particular meta-study or elsewhere (Huitema et al., 2011). Crucially, the researchers behind this meta-study did not assess formal and informal evaluation activities in the various jurisdictions they considered. In other words, the extent to which evaluations are placed to foster interactions between different governance centres in the EU has to date not been systematically studied. The existing literature by and large confines itself to cataloguing and analysing climate policy evaluation at broad brush, but not in fine enough detail to understand evaluation's role in interactions between governance centres (Haug et al., 2010; Huitema et al., 2011; Schoenefeld & Jordan, 2017). This constitutes a key gap in available knowledge in policy evaluation literatures.

3.6 Conclusion

So far, the literature reveals that the EU level, as well as Germany and the UK have been evaluation forerunners in the EU. However, the vast majority of evaluation literatures has focused on describing and to a certain degree assessing state evaluation institutions and actors in these countries. Very few studies have in turn focused on cataloguing or let alone assessing evaluation output at the EU level (but see Haug et al.,

2010; Huitema et al., 2011; Mastenbroek et al., 2016; Zwaan et al., 2016 for notable exceptions). And even less is known about the activities and outputs of informal evaluation actors. Thus, the current state of the literature is heavily lopsided, and potentially fails to address an important and vibrant aspect of evaluation activity in Europe. The following chapters address some of the gaps identified above.

Chapter 4 Methodology

4.1 Introduction

In seeking to apply the Ostroms' ideas about polycentric governance to climate policy evaluation, it is necessary to recognize that Elinor Ostrom and her collaborators did not only pursue novel theoretical endeavours. They also proposed a methodological approach that attempts to understand a world where, in her view, long-standing ontological and epistemological disputes have balkanized and at times paralyzed political science and related disciplines for far too long (E. Ostrom, 2006). In order to fully grasp her approach, it is thus necessary to not only explore its theoretical origins and developments, but also its philosophical foundations. In the first part of this chapter, I thus critically engage with the Ostroms' key ideas about political science and beyond, with a particular focus on how they respond to existing epistemological and methodological debates in Section 2 and Section 3. Building on these insights, Section 4 outlines how the current project seeks to implement these ideas, focusing on the main design and methods choices made in this thesis. This includes an overview of a database of climate policy evaluation studies (including a brief overview of the data contained therein), as well as an outline of the coding scheme devised for analysing policy evaluations from a polycentric perspective. Finally, this chapter discusses relevant ethical considerations, reflects critically on the research process, discusses limitations, and concludes.

4.2 The Ostrom approach to ontology and epistemology

The vast majority of ontological (how the world is) and epistemological (how we gain knowledge of the world) debates in political science and related disciplines in recent decades have boiled down to a fundamental clash between those who harbour

foundationalist/positivist beliefs (essentially the idea that the physical and social world exist independently of human conscience and interpretation) and those who argue in the anti-foundationalist/interpretivist tradition that the world is 'socially constructed', implying that reality is a product of the human mind and the interaction between different minds (Furlong & Marsh, 2010). Furthermore, it has been argued that each of these ontological traditions corresponds with epistemological and methodological approaches, with positivists preferring quantitative and 'objective' methods, and interpretivists stressing the need for qualitative, in-depth approaches to explore how humans construct the world around them. Symptomatically of these divisions, Furlong and Marsh (2010, p. 193) conclude that "[…] researchers cannot adopt one position at one time for one project and another on another occasion for a different project." According to these scholars, these positions are not interchangeable because they reflect fundamentally different approaches to what social science is and how it is conducted.

In recent years, however, there have been various attempts to soften the edges of these rather stylized and often entrenched philosophical fault lines. Or, in Elinor Ostrom's words, to make better use of the "the excessive energy devoted to factional fights" (E. Ostrom, 2006, p. 3). One such approach is critical realism, which accepts a positivist ontology while broadening its methodological reach to both quantitative and qualitative methods (Furlong & Marsh, 2010). However, this approach still retains many of the dualisms from the positivist/interpretivist divide. By contrast, Elinor Ostrom and colleagues have argued that the reasoning expressed by Furlong and Marsh (2010) and many others unnecessarily and unhelpfully balkanizes academic practice. In the spirit of making progress in this divided world, the 'Ostrom approach' is essentially a pragmatic one, focusing in its methodological variant more on research practice than theory, and avoiding "name calling" (E. Ostrom, 2006, p. 4). This approach is critically reviewed below, while engaging with various additional insights that have emerged from different quarters of political science and beyond.

The key underlying ontological proposition that Elinor Ostrom—along with many others in the social sciences—hold is that the social world is fundamentally different

from the biophysical world, and that this has implications for how we study the former. As she explains:

> The basic difference between the social world and the biophysical world is that the biophysical world exists whether or not humans reflect on it, but the social world is constituted by human thought, language, and action. Given the importance of language, a more serious threat to the future of our discipline than the lack of universal laws is our lack of common definitions for key terms we use including power, norms, and institutions. Can we ever escape from the "Tower of Babble" that we have created? (E. Ostrom, 2006, p. 4)

In other words, studying human beings and their social world is not the same as studying worms or molecules. This is because unlike worms or molecules, human beings are 'reflexive', meaning that they have at least the potential to react to the way scholars theorize about them (see Hay, 2002). Thus humans can "[…] contemplate, anticipate, and can work to change their social and materials environments […]" (George & Bennett, 2005, p. 129). In a reflexive social world where unwritten rules or social conventions exist only in people's minds and through commonly-shared knowledge, Elinor Ostrom holds that the quest for 'universal laws' is futile, because such 'laws' tend to hold only in highly constrained situations. In other words, related to the arguments developed in Chapter 2, context matters immensely (E. Ostrom, 2006; Sprague, 1982). Conceptualizing political actors as reflective agents means in turn that these actors are not immune to or insulated from insights generated by scholars engaged in theorizing and studying the social world (Hay, 2002).

This view is largely in line with other powerful criticisms of applying the foundationalist ontology in social science. As Furlong and Marsh (2010) explain, an influential critique of positivism suggested by Quine is that whatever knowledge we gain of the world passes necessarily though the filter of our senses, as well as our existing concepts. Consequently, our existing concepts and theoretical approaches extend our physical senses to perceive, prioritize, order, and categorize the bewildering amount of stimuli or 'data' that we can access on a particular question or phenomenon.

In the early 1980s, Elinor Ostrom contributed to related lines of reasoning by launching a critique of positivism in the form of an edited collection (E. Ostrom, 1982). As she explains there, "[t]o some extent the heavy emphasis on descriptive, empirical, quantitative work may have resulted from the naïve acceptance of a particular school of philosophy of science" (E. Ostrom, 2014a, p.214). However, as her later writings show, Ostrom would strongly disagree with the argument that holding certain ontological positions forces researchers to take particular epistemological stances, in contrast to Furlong and Marsh (2010). In part, this is because no matter how much researchers disagree about their philosophical convictions or dogmas, these considerations are usually only one consideration when choosing certain research methods —and perhaps not even the most important. As Poteete, Janssen, & Ostrom (2010, p. 10-11) argue:

> The influence of theory—and the implied influence of ontology—on methodological practice cannot be assumed and should not be over-stated. Theoretical changes can and do occur independently of changes in methodological practice […]. We argue that methodological choices are often driven as much by data availability or career incentives.

Related considerations led Elinor Ostrom and colleagues to a 'pragmatic' approach to research, which centres on the productive synergies of different research methods in practice, rather than their theoretical antagonisms – while however stressing that logical congruence between one's philosophical orientations and research methods remains important (E. Ostrom, 2006; Poteete et al., 2010). In essence, they argue that each methodological approach brings to the fore potentially useful elements—and thus advocate a pragmatic approach to research methodology, driven by the needs and constraints of researchers and their particular questions (Poteete et al., 2010). Facing increasingly complex systems that defy traditional reductionism, some philosophers of science have argued for a similar epistemological pluralism (see Mitchell, 2009). Taken together, even if higher-level ontological positions prove ultimately irreconcilable, productive synthesis may be possible through research practice and ultimately combining multiple approaches to look at a phenomenon. Much of the work of Elinor Ostrom and collaborators follows this line of thinking. The implications for this thesis

are thus to explain how the chosen method (i.e. database development and subsequent quantitative coding, see below) offers one important, but undeniably partial, view into the world of climate policy evaluation. It is partial in the sense that it cannot capture all of the relevant aspects of policy evaluation, but contributes crucially to developing a deeper understanding of this important governance practice.

While accepting the basic difference between the natural and the social world, in this thesis I draw on this pragmatism with a view to using the at least partially normative polycentric governance theory (see Chapter 1) to study empirical patterns of climate policy evaluation. I therefore focus less on defining an 'ideal', atomistic ontological and epistemological approach, but rather seek to maximize the congruence between the nature of my questions and the methods used to answer them, while remaining fully aware of the very real limits of this methodology (see Section 4.7).

4.3 Normative theory in social research: theory and empirics

While I argued in Chapters 1 and 2 that much of the Ostroms' work on polycentric governance contains undeniably normative elements, the idea that the social world differs markedly from the natural world goes a long way to reconcile tensions that emerge when seeking to reconcile a more positivist account of the social world with the use of normative theory in social research. In a nutshell, given that humans are reflexive, any theory about them or their behaviour has at least the potential to have a normative effect, because people may adjust their behaviour and choices in response to a theory, which can in turn become a self-fulfilling prophesy. For example, arguing that climate policy has become increasingly polycentric, and that polycentric governance can (or even should) yield positive results (e.g., E. Ostrom, 2010c; 2014b; Victor et al., 2005) may in turn affect the expectations of governance actors and thus their subsequent decisions and actions (as Chapter 1 argues, the 2015 Paris Agreement may be an example of this). In other words, governance theory on climate change and European integration has at least the potential to effect (normative) change in the (social) world (see Gravey, 2016). Building on the argument of the co-constructed nature of the

political world and reflexive political agents, *stricto sensu* the scholar becomes an additional political participant in constituting the political world. Based on the view explained above, inhabiting an a-theoretical, disinterested and uninvolved position—particularly with a view to something as normative as evaluating climate policy (see Chapter 1)—simply is not one of the available options. Therefore, at least in the social sciences, the scholarly enterprise always has at least an implicit normative character and may thus explain Elinor Ostrom's overt use of normative elements in her own work (see Chapter 2).

This does not mean, however, that empirical research is futile or even unhelpful. As political scientists and colleagues in other disciplines are well-aware, social structures are often slow to change, and they are influenced by a range of factors, including history, culture, and others (see Tilly & Goodin, 2006). Thus, while the 'Ostrom approach' holds that in principle theory should precede empirical investigation (Aligica & Sabetti, 2014a, p. 2), given that the social world is highly unlikely to change immediately and unequivocally in response to some theorist's propositions, there is room for a dialectic relationship between theory and empirical insights (E. Ostrom, 2014a). In other words, once launched, theory can enter into a co-productive relationship with empirical data, where theory stimulates data collection, and data in turn 'speaks back' to theory. It is in this spirit that I seek to subject some of the normative and positive claims regarding the role of information/evaluation in polycentric governance systems to empirical scrutiny (see Chapter 1). However, I do so with an awareness that such work also contributes to or at least engages with the larger normative project of polycentric governance and, crucially, with climate policy evaluation and broader debates on European integration and collaboration.

By and large, my approach aligns with recent developments in political science. While positivists once advocated strictly separating empirical investigation and normative theory, the advent of 'applied normative theory' has gradually led to intellectual cross-fertilization between what had become increasingly separate scholarly endeavours with often equally separate communities (Bauböck, 2008, p. 42). This trend has been particularly pronounced within scholarly communities working on policy

evaluation—where traditionally there has been a strong focus on normative aspects and at times very limited attention to empirical evidence (see Chapters 2 and 3). As Bauböck (2008) argues, subjecting normative theory to empirical evidence can add a sense of realism and reflection into normative debates and evaluate to what extent normative ideals appear workable in various political settings. Facing the argument that this amounts to policy advocacy, Bauböck (2008) suggests:

> Not to abandon normative theory altogether or to confine it to an arcane academic discourse, but to expose it instead to the full force of critique from explanatory theory and empirically grounded research to analyse the application context for normative ideas. (p. 59)

My approach in this thesis—empirically examining the emerging theory of polycentrism in the context of climate policy evaluation—follows the approach in the quote above. I thus accept to a degree that particularly in social and political settings a quest for 'value-free' research is indeed futile, but nevertheless endeavour to rigorously and systematically test the underlying values and assumptions of polycentrism. In doing so, I "[…] treat the empirical claims not as assumptions but as hypotheses […]" (D. F. Thompson, 2008, p. 498). Thus, once theory is launched, empirical evidence becomes an important 'helping hand' to its further development (D. F. Thompson, 2008, p. 500). Empirical research can thus make a very useful contribution to more normative debates on governance (see Smith, 2004).

Before I describe the specific research design used in the current project, a key point about generalization in the Ostrom tradition is warranted. One of Elinor Ostrom's key contributions is her more nuanced view of generalization than that which has been advocated in natural and social sciences. In a nutshell, her argument that context matters immensely in policy (see Chapter 2) renders problematic attempts to generalize from the individual characteristics of one governance context to another. Therefore, she argues that there are no "panaceas" (E. Ostrom et al., 2007). Yet in her work on natural resource governance, she argues strongly that much can be learnt from the large number of case studies that she analyses. However, her approach to 'generalization' or more general learning from case studies is grounded in the idea that direct generalization from

clearly measurable variables in natural resource governance (e.g., the number of fishers in a fishery, the size of their boats, the number of fish, the exact appropriation and decision rules etc.) is largely impossible, because particular contexts and large numbers of variables combine to produce unique governance outcomes that vary from one place to another (see E. Ostrom, 1990). This does not mean, however, that nothing more general can be learned from these cases. Elinor Ostrom proposes a focus on generalization at a higher level of abstraction, namely governance 'principles' rather than precise 'variables.' Her natural resource governance principles, drawn from multiple case studies, show higher-level, but not exactly/numerically specified guidelines that can be drawn from her cases (E. Ostrom, 1990; E. Ostrom, 2005). For example, she explains a need for low-cost, local conflict-resolution mechanisms or for continuous monitoring (E. Ostrom, 1990, p. 90). However, precisely how to design these mechanisms effectively depends on the particular situation. Taken together, Elinor Ostrom's approach to generalization thus clearly connects with her approach to conceptualizing analysis at different levels, that is, frameworks, theories and models (E. Ostrom, 2005, p. 27-29). Their assumptions and specificity vary, with frameworks being the most general approach (see Chapter 1). A generalization may not be possible at the same analytical level, but higher-level principles can be derived. This is how Elinor Ostrom conceptually reconciles the need for broader policy knowledge policy with the importance of context—an argument that matters immensely for the use of the knowledge that policy evaluations generate.

Policy evaluation literatures have also engaged in significant ontological and epistemological discussions, which matter for the factors considered at the centre of this thesis. For example, on the concept of context, Greene (2005) explains how for experimentalists and more quantitative evaluators, context is something to be controlled (i.e. a confounding, but not a relevant factor); realists see context as an additional explanatory factor (in addition to other factors) and finally, qualitative evaluators see context as inextricably bound up with the outcomes of a policy, thus requiring detailed description. In mirroring the debates in political science explained above, this thesis draws on Elinor Ostrom's approach as a way forward.

4.4 Research design and methods

4.4.1 Studying policy evaluation in polycentric systems

The way polycentric governance centres are thought to work guides the analyst's approach to studying them. One approach in traditional political science tends to place the state or large-scale institutions at the centre of analysis. In contrast, at the other end of the spectrum, behaviourists focus on individuals and seek to strip them of any external influence to expose the basic building blocks of human behaviour. The polycentric approach follows neither of these arguably extreme positions, but seeks to position itself somewhere in the middle in order to conceptualize individual and structural influences simultaneously. Most centrally, as Vincent Ostrom (1999b) [14] argues, the central unit of analysis becomes the individual, or in some instances groups of individuals, rather than 'government.' However, individuals are not 'atomistic:' Vincent Ostrom (1999b, p. 124) draws on key contextual variables to argue that

> […] the critical variables of concern to scholars in the polycentric tradition include (1) individuals; (2) decision rules; (3) sets of events; (4) outcomes; and (5) measures of performance.

In this thesis, I focus in particular on the fifth element, namely 'measures of performance,' here defined as policy evaluations—also in the context of the other elements, such as relationships between the individuals or groups in producing these evaluations (e.g., between formal and informal actors). Vincent Ostrom (1999b, p. 124) furthermore explains that there are multiple criteria against which we can evaluate policy outcomes. But

[14] Vincent and Elinor Ostrom were married, worked in the same research team and have significantly advanced polycentric governance theory over time (see Chapters 1 and 2). While Vincent Ostrom advanced key theoretical building blocks as early as the 1960s, Elinor Ostrom turned back to this work in the late 2000s in the context of climate change, while incorporating insights from her earlier work on common pool resources.

> [i]f evaluative criteria can be developed into general measures of performance, then different patterns of organization or different institutional arrangements can be measured in relation to common standards of measurement or yardsticks.

The latter can be understood as a call for 'harmonized' evaluation criteria. However, this view creates a key tension with arguments for more de-centralized policy evaluation detailed in Chapter 2. As a consequence, when studying the role of policy evaluation from a polycentric perspective, it is necessary to pay close attention to how this tension plays out in practice.

Following the pragmatic approach proposed by Elinor and Vincent Ostrom, this thesis uses methods that include both more quantitative approaches to understand the 'big picture' of key characteristics of the database of evaluation studies ('what is out there?'), but also more detailed, and in part qualitative, coding to understand a smaller number of evaluations and practices in detail. The following section unpacks this analysis, focusing on case selection, sampling, and coding, as well as a first descriptive overview of the evaluations that this thesis analyses.

4.4.2 Study focus and case selection

Research design refers to the overall conceptual architecture of a research project in light of the broader ontological and epistemological issues discussed above. Overall, the aim of this project is to build a theoretically and empirically informed account of climate policy evaluation practice in the EU from a polycentric climate governance perspective (see Chapter 1). To do so, I focus on the EU level—the main locus of climate policy development and evaluation (see Chapter 1), as well as on Germany and on the United Kingdom (UK), which are the EU's two largest economies and carbon dioxide emitters, and thus especially relevant for climate change, and who are also active climate policy adopters and evaluators. This makes the EU a 'crucial case' (Eckstein, 2000) or a 'critical case' (Flyvbjerg, 2006) for the role of policy evaluation in polycentric governance systems. As Eckstein (2000, p. 148) explains, a crucial case is

one "[…] that *must closely fit* a theory if one is to have confidence in the theory's validity, or, conversely, *must not fit* equally well any rule contrary to that proposed" (emphasis in original). As Chapter 1 explains, starting from the point that European Union climate governance is polycentric *par excellence* (E. Ostrom, 2010c; E. Ostrom, 2014b; T. Rayner & Jordan, 2013), makes this a 'most likely' case for an active role of evaluation in the emergence of a polycentric climate governance system. In other words, if evaluation does not facilitate polycentric governance in the case of EU climate change governance, it is highly unlikely to do so in other polycentric arrangements. In addition to these theoretical considerations, I am a German native speaker, which allowed me to analyse evaluation studies in the German and English languages. In order to capture the full range of climate policy evaluation activities in Germany, in the UK and at the EU level, I considered both formal (state-driven) and informal (society-driven) evaluations (see Chapter 2).

Overall, my goal was thus to locate and catalogue climate policy evaluation studies ('evaluations' from now on) generated up to 2014,[15] a time period during which the EU substantially increased its climate policy-making (see Jordan et al., 2010). Earlier studies under the auspices of the Adaptation and Mitigation Strategies (ADAM) Project (see Hulme, Neufeldt, Colyer, & Ritchie, 2009) analysed *ex-post* climate policy evaluation activities in the EU between 1998[16] and March 2007 (Haug et al., 2010; Huitema et al., 2011). Since 2007, there has been no systematic audit of evaluation activities, even though there have been discussions on the need for more meta-analyses of climate policy evaluations (see Wörlen, 2011). The current study extended these analyses to 2014 while including a much wider range of evaluation actors, considering a vital area of EU climate policy development, and applying a novel theoretical perspective.

---

[15] In line with Haug et al. (2010), the goal was to identify and catalogue as many evaluations as possible, given time and resource constraints. As I explain below, this is likely to have generated an extensive, but probably not entirely exhaustive, list of evaluations.

[16] Prior to 1998, there were very few *ex-post* evaluation studies (see Huitema et al., 2011).

4.4.3 Assembling the evaluation database

The first step of this research was to build a novel database of *ex-post* climate policy evaluations published between 1997[17] and 2014. The focus is on published evaluations. Recognizing that evaluation is a wider process involving a range of actors and many different types of interactions (e.g., participation of various actors), this study focuses on the concrete and publically available—but inevitably only partial—outputs of that process (see Chapter 1). Researchers who had worked on the ADAM project (see above) kindly provided their database of evaluations, but they used a slightly different operational definition of *ex-post* evaluation studies than I did for my own database. Specifically, the ADAM project included articles from professional academic journals, as well as book chapters, whose primary aim may not have been to analyse existing policy with a view to establishing its performance or worth (the standard definition of an evaluation – see Chapter 1) and often providing recommendations, but rather collecting data and/or testing theory developed in academia or otherwise.

In contrast to this broad definition of evaluations, and in line with the definitions of evaluation discussed in Chapter 1, for the current project I chose a narrower definition of *ex-post* evaluation and therefore excluded publications in academic journals while including a wider range of organizations and actors given the broad orientation of polycentric governance with a keen interest in formal and informal actors (see Chapter 2). The operational definition concentrates on four key dimensions, namely the focus, the purpose, the analysis, and the temporal orientation of an evaluation:

- ***Focus:*** Evaluations of 'public policy' only, that is, policies put forward by governments/governmental actors at the EU level, as well as at the national level in Germany or in the UK. This is in line with Vedung's classic (1997) definition of evaluation (Chapter 1). My database therefore does not include evaluations of

[17] The general search extended to the middle of the twentieth century, but the first evaluation located was from 1997. This corroborates Huitema et al. (2011), who also pointed to the extremely low numbers of climate policy evaluations before the mid-1990s.

other governance approaches, such as city networks. In substantial policy terms, and following Huitema et al. (2011), I focused on evaluations of policies that ultimately seek to reduce greenhouse gas emissions and which countries routinely report to the United Nations (UN). I limited the search to mitigation policies; that is, policies that aim to reduce greenhouse gas emissions. The database therefore excludes all evaluations on climate adaptation policy (i.e. policies that seek to enable societies to live with certain levels of climate change). I only included evaluations of policies in all sectors (such as transport, energy or agriculture) that address greenhouse gases (see above). A policy sector may be defined as "a place where an empirically observable set of actors defines a general set of rules and norms for commonly accepted characterizations of policy-relevant issues or concerns" (J. Rayner, Howlett, Wilson, Cashore, & Hoberg, 2001, p. 320). To ensure a policy focus on climate change, in each case I keyword searched the evaluation for 'climate change' and 'greenhouse gases'[18] to determine whether or not the evaluation focused on a policy that was thought to have a significant climate change impact. Geographically, the database only contains evaluations that address policies enacted at the EU level, or at the national level in Germany or in the UK, that is, evaluations of regional or local policies were not included. Furthermore, the database only includes evaluations that were funded by formal and informal actors at EU level, in Germany or the UK (see below for a more detailed discussion of the 'funding' criterion). Evaluation 'funding' here is a broad category that includes both an organization funding its own evaluation, as well as an organization funding others to conduct the evaluation (a more restrictive category would be commissioning evaluation, which means that one actor commissions another to conduct the evaluation). For example, a policy evaluation of a UK policy funded by actors in the UK would be included in the database (and analogously for the EU level and Germany). Furthermore, a German or UK policy evaluation funded by EU level evaluators

[18] In the German-language evaluations, I used the German equivalents, namely "Klimawandel" and "Treibhausgase."

would also be included in the database, as would evaluations of EU level policy funded by German or UK-based evaluators. By contrast, evaluations of EU level policy funded by for example US-American actors, or evaluations of German/UK/EU policy funded by Belgian evaluators would not be included in the database. However, evaluations funded by actors at the EU level, in Germany or in the UK (conducted anywhere) were included.

- ***Purpose:*** The database only includes evaluations whose primary goal is to evaluate a policy and provide some type of recommendations or policy-relevant conclusions (i.e. not academic theory-testing or use as a case study in broader theoretical arguments), as has been argued in relevant evaluation literatures (see Chapter 1). Because publications in academic journals are not included, my operational definition is narrower than that used by researchers in the ADAM project (see Haug et al., 2010; Huitema et al., 2011).
- ***Analysis:*** To be included in my database, evaluations needed to include sufficient and systematic analysis – containing and analysing novel information collected for the evaluation, or alternatively recombining existing information and arguments in new ways. Therefore, short documents such as press releases/policy briefs or political position statements were not included (see Haug et al. (2010), who adopted a very similar approach). However, if a policy brief/position statement was based on a more extensive evaluation that satisfies these criteria, I sought to locate the original evaluation for inclusion in my database.
- ***Temporal orientation.*** The main orientation of the evaluation had to be *ex-post*, that is, retrospectively looking at the outcomes/performance/worth of an existing or terminated policy. Evaluations mainly anticipating future effects (*ex-ante* evaluations or 'impact assessments' – of which there are many (Turnpenny et al., 2016) of some proposed policy measure were not included in the database – again following the approach by Haug et al., (2010).

Table 4.1 summarises these inclusion and exclusion criteria based on a combination of evaluation funding and the policy level.

Table 4.1: Evaluations included in the database

Evaluation funders from… \ *Policy level*	DE national	UK national	EU	Regional/local (anywhere)	All else
DE	✓	✓	✓	×	×
UK	✓	✓	✓	×	×
EU level	✓	✓	✓	×	×
All else	×	×	×	×	×

Overall, I searched for evaluations with the aforementioned geographical orientation published any time until 31 December 2014. I only included evaluations in my database that were at least in principle in the public domain; that is, they did not require special access permission arrangements and were available in electronic format on the Internet (or as part of an existing database) at the time of searching (2014-2016). For example, I did not file Freedom of Information (FOI) requests in order to retrieve evaluations. This approach is in line with earlier research in this area (e.g., Huitema et al., 2011). I focused only on public evaluations, because all but the most well-connected actors in a polycentric governance system would be unlikely to know of or be able to request/use evaluations that are not publically available. Furthermore, in practical terms there is currently no systematic way of knowing which non-public evaluations exist, thus it would not have been possible to assemble a reliable population of these evaluations (and the totality of climate policy evaluations outside those publically available remains unknown). Last, earlier literature suggests that specific aspects of freedom of information differ in Germany, in the UK and at EU level so that conducting related requests may have yielded different results as a function of different underlying approaches (e.g., Bugdahn, 2008).

Aside from drawing upon the database from the ADAM project, I located evaluations through web searches by drawing both on my personal knowledge of evaluation organizations, using key evaluation databases where available (see Table 4.2 below), and using ADAM project evaluations to identify key organizations that have

conducted policy evaluations before. In addition, I attended several relevant conferences and workshops with practitioners (see the acknowledgements for a list), where I asked for advice on how to locate publically available evaluation sources through 'snowballing.' In each case, I recorded where I first encountered an evaluation in order to build a knowledge base on where most of the evaluations could be found.

Given that there is no central European database or source for climate evaluations that encompasses all jurisdictions of interest in this study, and especially because nobody systematically collected informal evaluations anywhere, I had to build a new database from various sources. I began assembling my unparalleled database by combining relevant climate policy evaluations from a range of existing databases, which were mainly compiled by academics in the context of other projects. Furthermore, I collected evaluations from individual organizational websites online. Table 4.2 provides an overview and brief description of these sources.

Table 4.2: Source databases

Database title	**Key characteristics**	**Source**	**Number of evaluations retrieved**
ADAM project database	• Contains 259 *ex-post* climate policy evaluations. • Includes published academic studies as evaluations.	Huitema et al. (2011) and personal contact with researchers involved in the project.	130
Monitoring Mechanism reports to the European Environment Agency	• The Monitoring Mechanism is a tool to collect mainly *ex-ante* data on climate policies across the European Union. • Reviewed reports published in 2009, 2011, and 2013 for additional evaluations.	European Environment Agency (EIONET website and personal contact with EEA officials).	14
Forschungsradar Energiewende Database	• Large, online database of evaluations on the German Energiewende.	Online at: http://www.forschungsradar.de/studiendatenbank.html.	127

	• Operated by the German Agency for Renewable Energy and co-funded by the German Federal Government.		
Warren Demand-Side Management Database	• Large meta-analysis by Peter Warren at the UCL Energy Institute. • 200+ evaluations.	The database is published in Warren (2014)	1
European Commission Smart Regulation/Evaluation Database	• Online database with evaluations conducted by the European Commission.	Available online at: http://ec.europa.eu/smart-regulation/evaluation/search/search.do.	18
Mastenbroek et al. (2015) dataset	• Dataset with 216 evaluation reports from 2000-2012 compiled from various sources.	See Mastenbroek et al. (2015). Stijn van Voorst kindly provided information on a number of relevant climate policy evaluations via email.	6
European Commission Multi-Annual Overviews of Evaluations & Impact Assessments	• Reviewed all Commission Multi-Annual Overviews between 2002 and 2009.	Available as official publications from the European Commission.	2
Eureval Database	• Dataset from private consulting company with 144 evaluations in total.	Thomas Delahais, who was involved with building the database, kindly gave permission to use it.	1
EU Climate Policy Bibliography (EUI)	• Online database with mainly academic climate policy evaluations.	Available online at: https://cprubibliography.wordpress.com/.	3
N/A (specific, individual collection for this thesis)	• Individual organizational websites and other internet sources	Various (see Appendix 2 for the full list).	316
My database (for comparison)	• See above.	See above.	618

The research involved reviewing each database in detail in order to extract relevant evaluations for the current project and subsequently encode them in my own database. Following this database review, I browsed organizational websites to locate additional climate policy evaluations. I started with key organizations that had emerged in the database review above and checked their individual websites for additional evaluations. Overall, this process was lengthy, sometimes difficult, and cumbersome. It required searching extensively for evaluations on individual websites, and reviewing a vast amount of evaluations in order to identify those that matched my selection criteria (see above). Following initial review, I presented the list of organizations I had consulted three experts[19] with relevant experience in this field, who checked for completeness and suggested additional organizations to consider. For a complete list of sources and the activities used to locate them, see Appendix 2. Overall, generating the database took over two years (2014-2016). In contrast to other studies, such as in the impact assessment domain (e.g., Fritsch et al., 2013), where governments have set up databases that can be combined by researchers, the source databases I used ultimately returned about half (48.87%) of the evaluations in my database (see the last column in Table 4.2 for the complete breakdown), and thus necessitated a wide-ranging and time-consuming search on individual organizational websites. This approach was particularly necessary for 'informal' evaluations (i.e. evaluations funded by societal actors, see Chapters 1 and 2), which have so far not been catalogued in an existing database after 2007 (see Huitema et al., 2011). Taken together, this generated an extensive and novel database of 618 climate policy evaluation studies in the UK, Germany and the EU. Novel because it covers both a far greater range of actors than had previously been considered and a period of rapid climate policy development in the EU. Extensive because it is nearly three times as large as the ADAM database, and because I only ended the search for evaluations once all the sources I and experts in the field could identify had been reviewed.

[19] Prof Andy Jordan, Mr. Christoph Priebe, Dr Tim Rayner.

4.4.4 Database overview

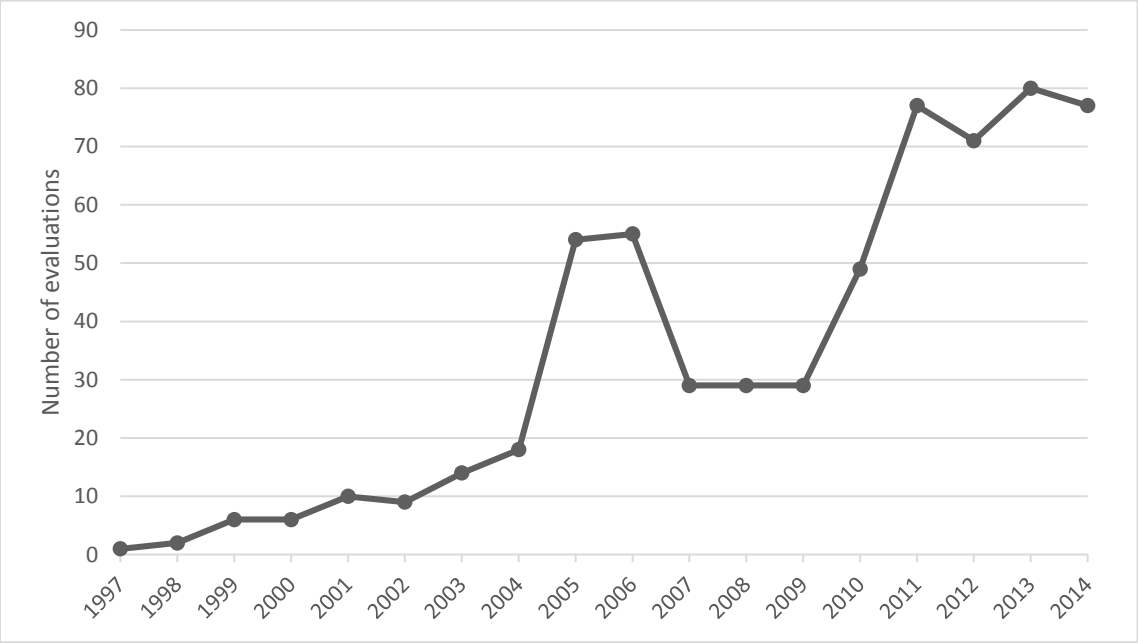
General overview. The first stage of the analysis involved creating a set of overview statistics on the whole database (N = 618). To this end, I extracted information on a small set of criteria from the evaluations and where necessary from the websites that published them, including the publication year, who conducted the evaluation (hereafter, 'the evaluator'), including broad evaluator categories (e.g., research institute/university, private consultancy, government body etc.), and who paid for the evaluation (the evaluation funder), including broad funder categories (e.g., policy makers, foundations, or environmental groups). Furthermore, I catalogued the climate policy sub-area that the evaluation addressed (e.g., renewables, transport, waste or cross-cutting). The analysis revealed some challenges. First, it was not possible to determine the evaluation funder for 74 evaluations, which translates into 11.97% of the entire database. These evaluations remained in the database, but received the qualification 'Not known' for this category.

Second, when multiple organizations from multiple governance centres funded evaluations together, I used the label 'EU' as the country label, indicating that they came from more than one EU country and to reflect the fact that the evaluation emerged from joint efforts of funders in several EU countries. When multiple types of organizations were involved in conducting an evaluation, I used the category of the first, or lead, evaluator for the evaluator category. For example, if an evaluation was led by a research institute with contributions from a commercial consultant, I coded the evaluation as 'research institute.' The following section contains a description of these general overview statistics on all evaluations contained in the database. Understanding the database in general terms is a necessary precondition in order to choose a smaller sub-set of evaluations for further analysis (see Chapters 5, 6, and 7).

The data collection yielded 618 evaluations focusing on climate policy at the EU level, or in Germany or in the UK (both national-level policy only) between 1997 and 2014. Figure 4.1 reveals that while the number of evaluations per year grows until the mid-2000s, there are notably fewer evaluations 2007-2009, when evaluation output drops to under 30 evaluations per year, in comparison to 57 evaluations produced in

2006 (a nearly 50% drop). However, by 2010 the evaluation output resumes previous levels and begin to surpass them in the following years. After 2011, the number of yearly evaluations remains at a high level with minor fluctuations with about 70-80 evaluations per year. Thus, these data are generally in line with Huitema et al. (2011), who detected strong climate policy evaluation growth until 2007 (although, notably, their sampling frame is somewhat different).

Figure 4.1: Climate policy evaluations over time: EU level, DE, UK (N = 618)

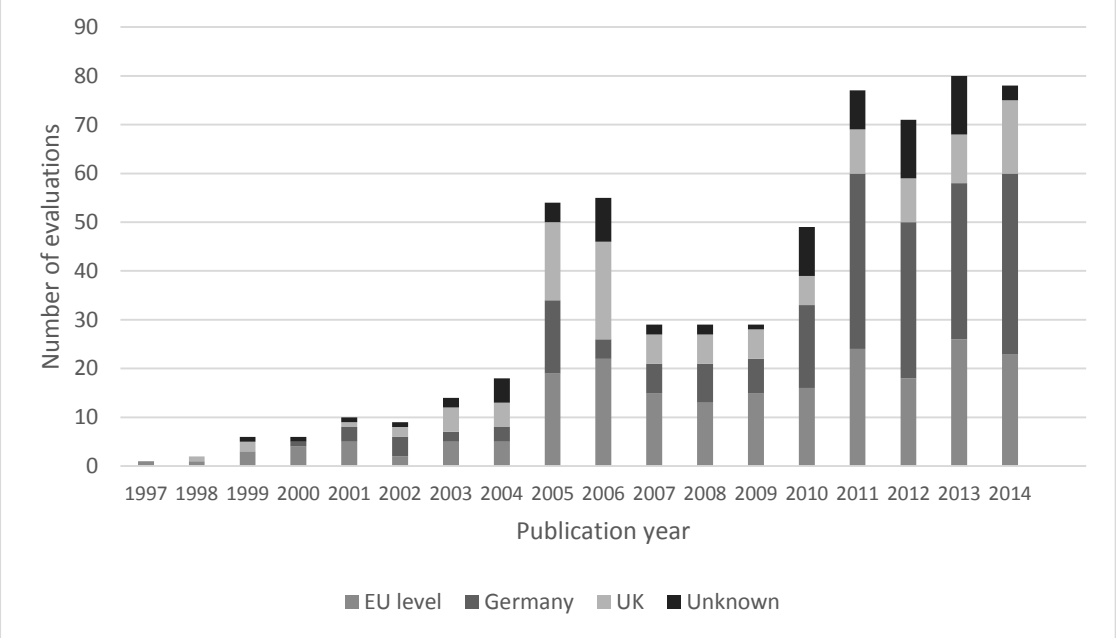


In order to gain a deeper understanding of these emerging patters and their potential explanations (see Chapter 8), a crucial step is to further disaggregate these data in order to understand (1) who has funded the evaluations, (2) who has conducted the evaluations, and (3) on which governance centres they focused. While previous studies have pointed to relatively high numbers of commissioned evaluations (Huitema et al., 2011), they did not yet distinguish between the location of those who fund and those who conduct evaluations—they focused mainly on the jurisdiction that evaluations were

concerned with; see Haug et al., (2010).[20] The following sections provide the overview data on these three dimensions with a view to further unpacking the data presented in Figure 4.1.

Evaluation funders. The location of the evaluation funders is a good proxy for locating the original 'impetus' of evaluation. As Chapters 2 and 3 discussed, this is relevant because producing evaluations requires significant resources. Doing so provides a much more precise set of data on the origin of evaluation. Figure 4.2 thus presents the number of evaluations over time against the location of the evaluation funder.

Figure 4.2: Evaluations funded at the EU level and in DE & UK over time (N = 617)[21]



While the overall growth in evaluations is reflected here, Figure 4.2 reveals that funders in the three governance centres analysed here do not drive this growth evenly.

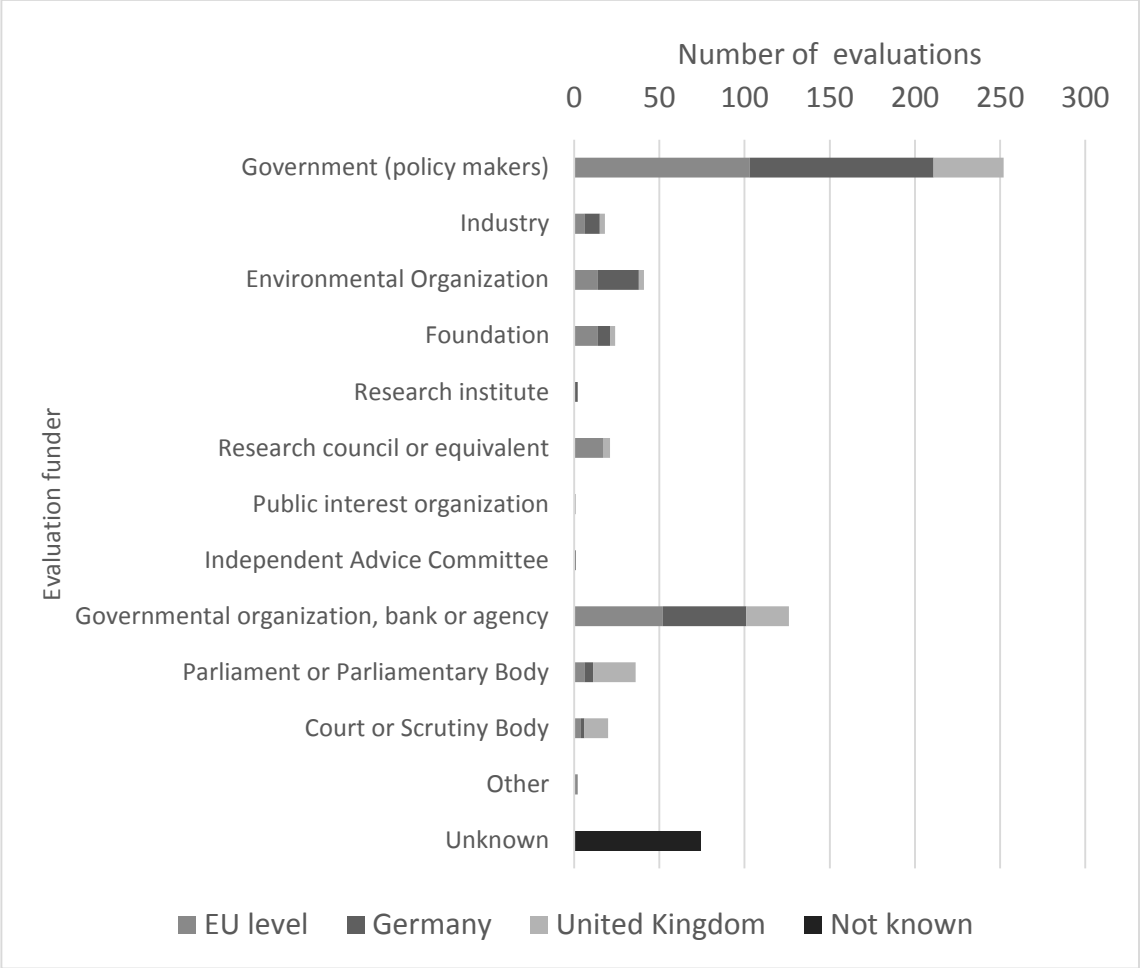
[20] These authors relied on double-counting when a report concerned more than one governance centre.

[21] One evaluation did not include a publication date and is thus excluded here.

While formal and informal actors in the UK and in the EU funded a comparatively high number of evaluations published in 2005 and 2006 (with actors in Germany funding comparable levels in 2005, but much less in 2006), the number of evaluations funded by actors in the UK grew at a much lower rate than in Germany after 2010. Actors in Germany funded smaller numbers of evaluations early in the period studied (2005 is somewhat of an exception), but starting in 2010, they surpassed all others, including EU level actors, in funding evaluations. The data thus reveal a climate policy evaluation 'funding boom' in Germany starting in 2010. It should also be noted again that it was not possible to decipher who funded 11.97% of all evaluations (74 evaluations across all years).

Crucial for this thesis is the distinction between formal (state-led) and informal (society-led) evaluations (see Chapter 2 and below). Taking again a funding 'impetus' perspective, Figure 4.3 shows that in the different governance centres, different types of organizations fund evaluation. Notably, courts or scrutiny bodies, as well as parliaments, play an important role in funding climate policy evaluations in the UK, whereas there is comparatively more involvement of environmental organizations, state-owned banks, governmental agencies, and policy-makers themselves in Germany and at the EU level. Strikingly, as far as evaluation funding goes, the types of evaluation funders in Germany and the EU appear to more or less resemble each other whereas the UK differs notably.

Figure 4.3: Evaluation funders by organizational type and location (N = 618)



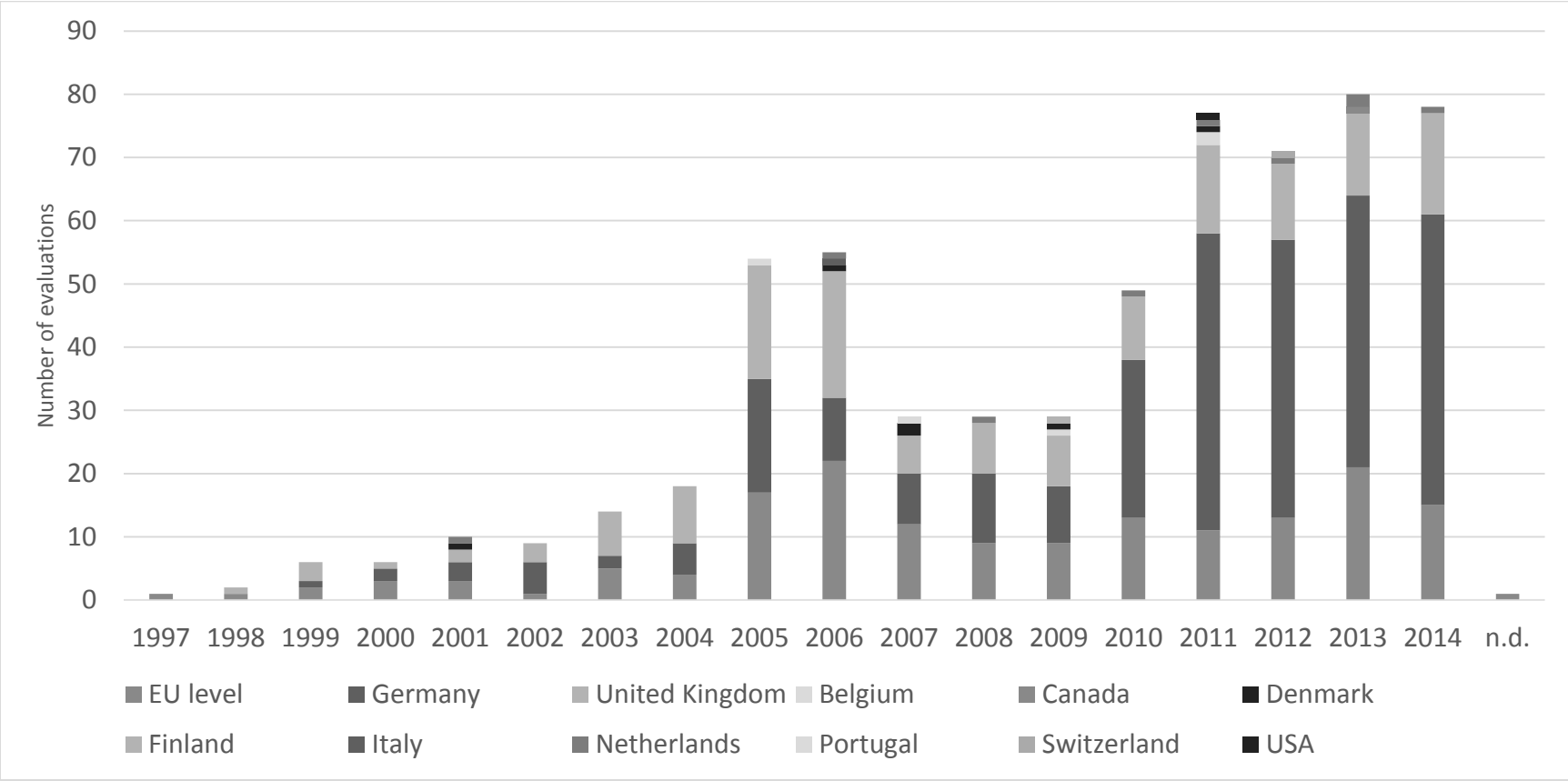
For the purposes of this thesis, all evaluation funders who are linked with the state (i.e. part of government or drawing on public resources) will be considered 'formal.' This is in line with the relatively broad definition of 'the state' put forward in Chapter 2. However, as Chapter 8 explains further, it is important to recognize that, as Figure 4.3 reveals, the 'formal' category contains a range of organizational types with often very different functions within the state. For example, the Departments of the UK government have a very different role from the UK Parliament or the UK National Audit Office (NAO), but crucially, all draw on public resources to operate. Drawing on the categorization here, the 'formal' category thus includes courts and scrutiny bodies, parliaments, governmental organizations, state-owned banks, governmental agencies,

independent advice committees, research councils, research institutes (which often either draw direct monies from public actors or significantly depend on the public purse for contracts), as well as (executive) government. In total, formal actors funded 458 evaluations, or 74.11% of the database.[22] By contrast, informal actors that funded evaluations included industry, environmental pressure groups, and foundations or public interest organizations. Compared to the formal category, informal actors funded a relatively limited number of evaluations (84 evaluations or 13.59% of the overall database). Thus, formal, state-linked actors were the main financial supporters for the climate policy evaluations in the database in all three governance centres.

Evaluators. The second key perspective on evaluation is the geographical location of the evaluators. Given the previously noted practice of funding others to conduct evaluation studies, evaluation funders and evaluators are not necessarily located in the same governance centre (but self-funding is possible). Figure 4.4 thus presents the distribution of evaluators across governance centres and time.

[22] The complete breakdown by evaluation funders is as follows: Total N = 618; of which formal evaluations = 458, informal evaluation = 84; unknown = 74; organizational type 'other' = 2.

Figure 4.4: Evaluations by the location of the evaluators over time (N = 618)[23]



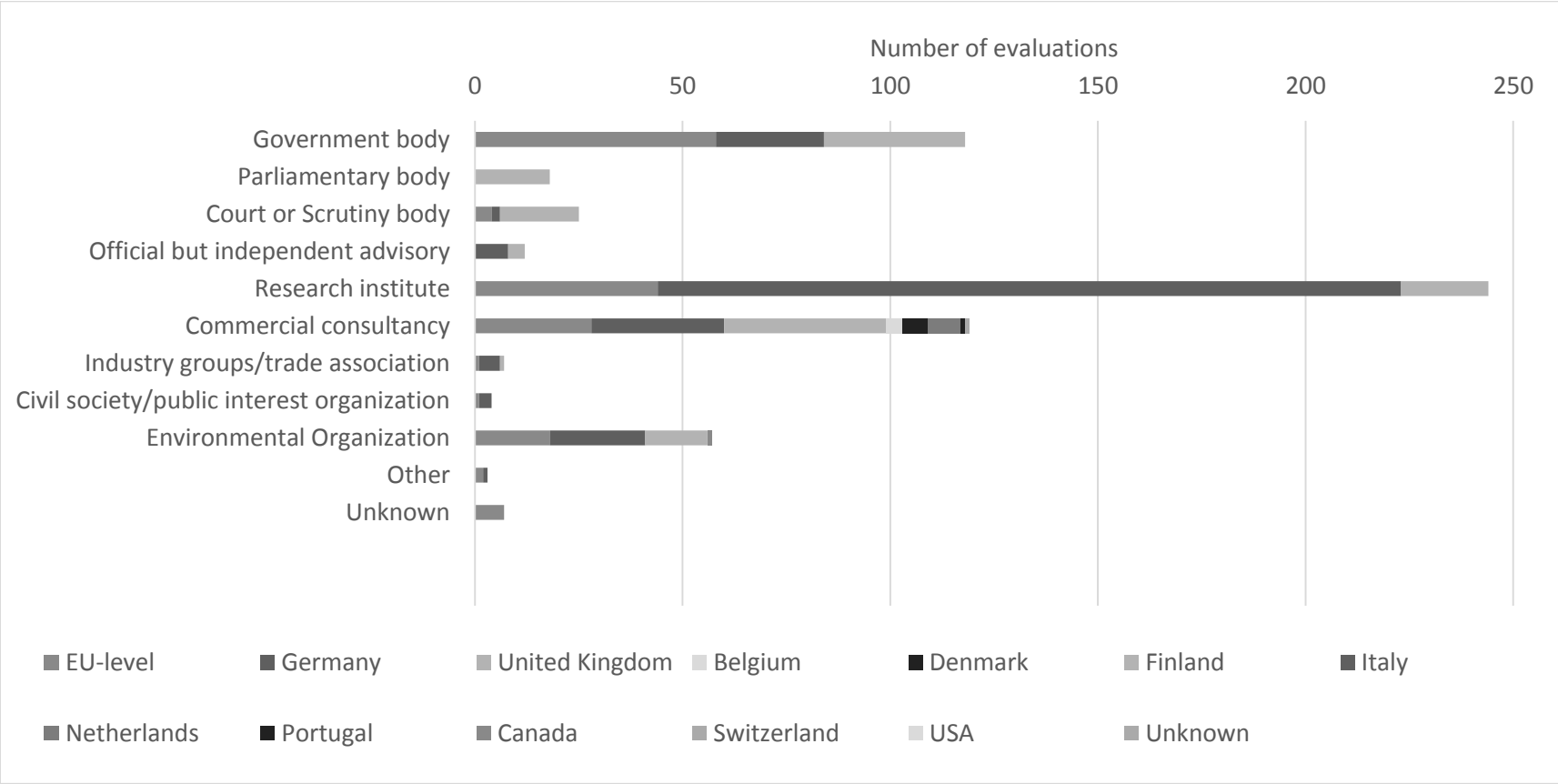
[23] The publication year could not be determined for one evaluation (indicated as n.d.).

Figure 4.4 reveals that across all years, the majority of evaluations were produced by actors at the EU level, in Germany, or in the UK (the largest sections in the bars for each year). The number of evaluations that evaluators from each governance centre produced varies over time. In terms of the location of evaluators, the EU level, Germany, and the UK by and large resembled each other until 2009. From 2010, compared to the EU and the UK, which remained similar, evaluators in Germany conducted nearly two to three times as many evaluations between 2010 and 2014. Thus, in addition to many of the climate policy evaluations being funded in Germany (see above), many evaluations were also being conducted in Germany. As explained in the previous chapter, it should be noted that these data reflect the geographic location of the 'lead evaluator' only, as many studies were conducted by consortia made up of multiple organizations.

Another way of looking at the same data is to consider the organizational type of the evaluators across the whole database (see Figure 4.5). These data were obtained by reviewing the author information in each evaluation and, where necessary, on the source website or similar. The categories were obtained from Huitema et al. (2011) and adapted to suit this database. Figure 4.5 shows that research institutes conducted the largest share of evaluations (just under half of the evaluations in the database), followed by commercial consultancies and government bodies. The low number of evaluations conducted by civil society organizations and industry/trade associations is also notable. Environmental organizations conducted just under 60 evaluations. Furthermore, Figure 4.5 reveals that German research institutes produced a very sizeable number of evaluations.[24] In the other categories, evaluators in these three governance centres produced comparable numbers of evaluations, with the exception of government bodies, which appeared to be more active in climate policy evaluation at the EU level compared to Germany and the UK.

[24] The category 'research institutes' includes universities; however, the number of evaluations conducted by universities is very small.

Figure 4.5: Evaluator category by evaluator country (N = 617)[25]



[25] See above – the organizational category of the evaluator could not be determined in one case.

Geographical focus of the evaluation. The third perspective in this general overview is to consider the governance centre on which an evaluation focuses. This perspective thus analyses where evaluations are directed. Figure 4.6 summarizes the number of evaluations focusing on climate policy in each of the three governance centres.

Figure 4.6: Evaluations of climate policy at the EU level, DE & UK over time (N = 617)[26]

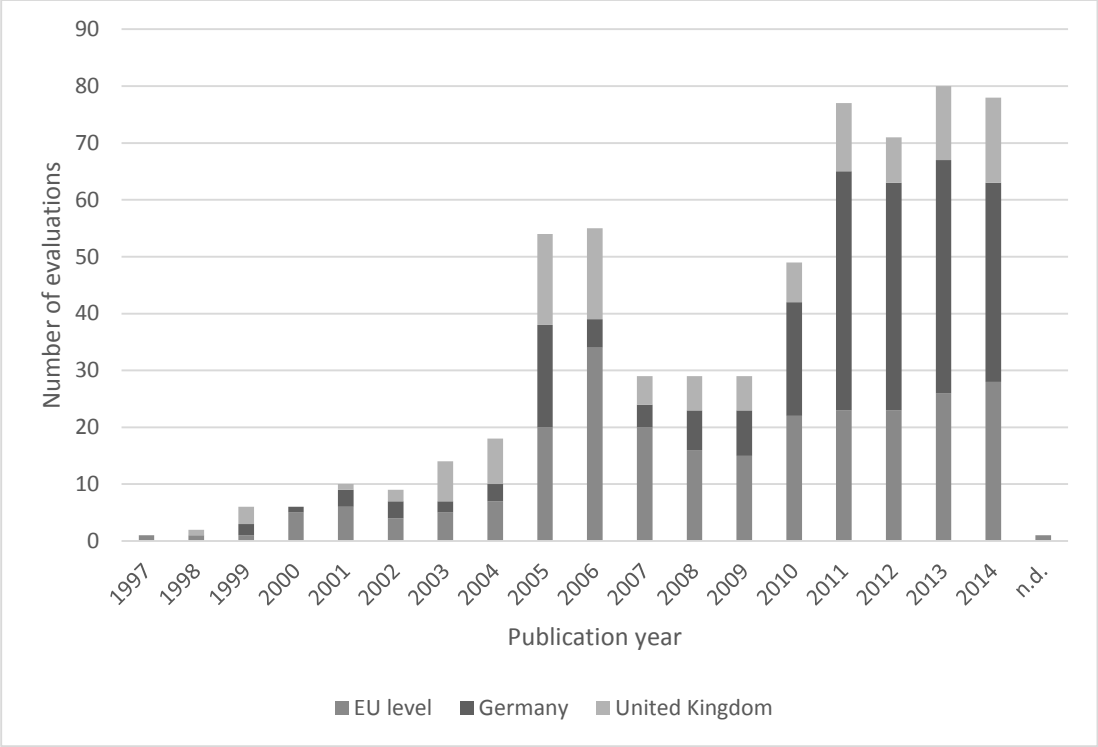
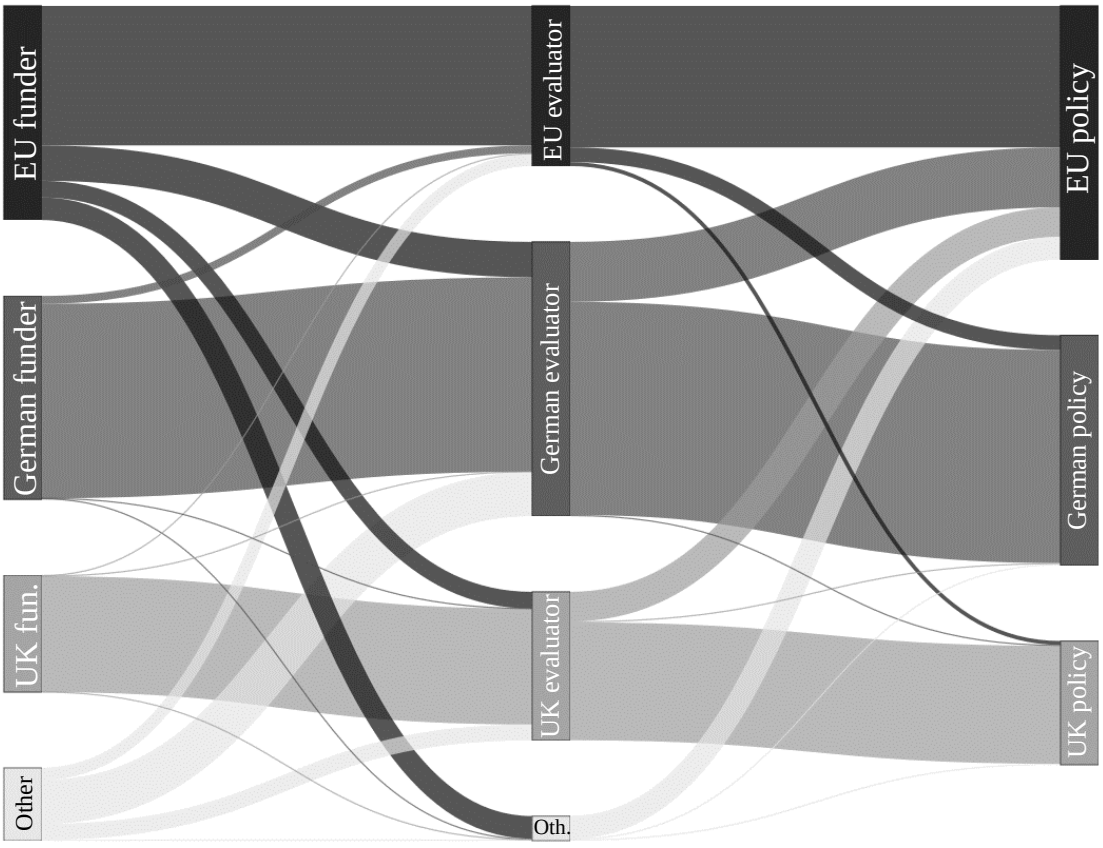


Figure 4.6 reveals that the number of evaluations focusing on climate policy in each governance centre fluctuates over time. While early on there was less evaluation of climate policies in Germany, this reversed from 2008 onwards, when German climate policy became the most evaluated dimension compared to the UK and the EU level from 2011 onwards.

[26] In one case, the evaluation year could not be determined; see above.

Figure 4.7 then draws together the aforementioned information on the number of evaluations by funder, evaluator, as well as the location of the policy under evaluation. The thickness of the connectors between funders, evaluators, and the evaluation in Figure 4.7 represents the number of evaluations with the respective characteristics (i.e. the bigger the number of evaluations, the thicker the connector). It demonstrates a strong congruence between the location of the funder, the evaluator, as well as the policy under evaluation. By and large, German funders tend to fund German evaluators to evaluate German policies, and so on for the EU level and the UK.

Figure 4.7: Evaluation funders, evaluators and policy under evaluation (N = 618)



Note: the thickness of the connectors represents the number of evaluations with the respective characteristics.

Climate policy sub-types. The fourth key dimension to understand the evaluation output is to consider on which climate policy sub-type each evaluation focuses. Figure 4.8 thus presents the number of evaluations by country and climate policy sub-type.

Figure 4.8: Evaluations at the EU level, DE & UK by climate policy sub-type (N = 618)

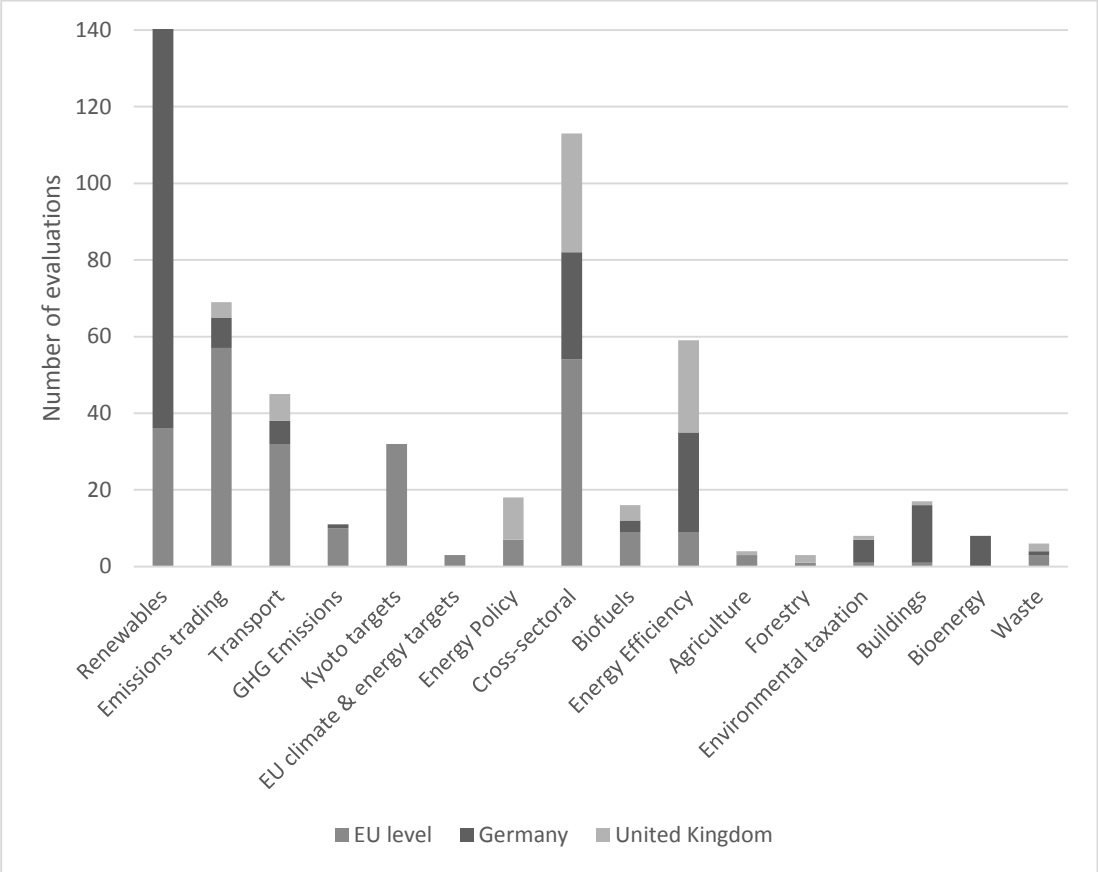


Figure 4.8 reveals that evaluations on climate policy at the EU level and in Germany and the UK differ in their focus on the climate policy sub-type. For evaluations focusing on Germany, the large number of evaluations on renewables policy stands out, but also evaluations on environmental taxation, buildings and bioenergy. In many ways, this is not surprising, given Germany's strong policy preference for renewable energy through the *Energiewende* and related efforts to boost energy efficiency. Historically, environmental taxation has also played an

important role in Germany since the early 1990s. Evaluations on emissions trading, transport, greenhouse gas emissions and various targets (Kyoto and EU) focused mainly at the EU level. Again, this is unsurprising given the central role the EU plays in the EU Emissions Trading System, as well as in assessing policy achievements against international climate targets. However, notably on energy efficiency, most evaluations focused on Germany and the UK, with relatively few evaluations conducted on policies in the EU as a whole. Furthermore, given the EU's tendency for package programs and the UK's 'technology neutral' attempts at making climate policy, the number of cross-sectoral climate policy evaluations appears correspondingly greater in the UK and at EU level than in Germany.

4.4.5 Analysis with a novel coding scheme

In order to gain a deeper understanding of the evaluations in the database, and study the key aspects discussed in Chapter 2, I developed a novel coding scheme in order to analyse a sub-set of the evaluations (for the results, see Chapters 5-7). Starting from the long-standing standard evaluation categories reviewed in Chapter 2 (such as time, formal/informal evaluation, etc.), I worked backwards to the foundational ideas of polycentric governance (Chapter 2)—namely self-organization, context and interacting governance centres—in order to turn these into empirically usable coding categories. Doing so involved an iterative process, where I first developed a draft coding scheme, reviewed that draft with experts in the field[27] and made further improvements. Once I arrived at a workable draft, I conducted pilot coding to test the usability of the scheme, and made adjustments whenever I noticed aspects that required further work.

The coding scheme used a mixture of deductive and inductive coding in that it started from a number of coding categories that have proven relevant in evaluation literatures and that are linked to the previous discussion on polycentric governance (see Chapter 2). For example, drawing on the discussion on context, the coding scheme sought to capture the length of time over which the climate policy is assessed

[27] Professors Mikael Hildén, Dave Huitema, and Andy Jordan.

or the extent to which the evaluation pays attention to the political environment in which a policy operated. However, I also kept the coding scheme open, allowing for flexibility to generate additional categories that emerged in the coding process. The aim was to generate a broad overview of the way in which context has been taken into account (or not) in climate policy evaluations, starting with standard evaluation categories, but not being limited by them. In this iterative process, I added the category of 'reference to climate science' in the contextual elements, which is quite specific and relevant to climate policy. Whenever I encoded a particular score, I recorded the relevant page number and justification for that score.

As a final check before beginning the analysis, I tested the coding scheme with a colleague[28] in order to assess the inter-rater reliability of this analysis tool. As D'Lane, Love, and Sell (2012, p. 42) detail, "[r]eliability in coding means that the biases inherent in the observers/researchers are substantially less than the 'true variation' of the behaviour being coded." In practice, we both coded three evaluations independently and then compared the results. Initially, we on average agreed on 53% of the items across the three evaluations (the scores were 64%, 50% and 46%). However, in the vast majority where we did not have immediate agreement, the scores varied by a point or so on a five-point scale. Following the independent coding, we discussed each score in detail and were in this process able to agree on a final score, as well as to calibrate the coding scheme in order to clarify it. Following the calibration of the coding scheme, I analysed the remainder of the sample (see below) alone, given time and resource constraints. Table 4.3 summarizes the key elements of the coding scheme. See Appendix 3 for the full coding scheme including the full scales.

[28] I am grateful to Christoph Priebe for his generous help.

Table 4.3: Summary of the coding scheme

Foundational idea	**Item**	**Scale**
Self-organization	Stated or implied purpose of the evaluation (Categorical)	Legal requirement, accountability, learning, none/not clear.
	Stated or implied target audience of the evaluation (Categorical; multiple mentions possible)	Politicians, bureaucrats, wider public, not clear.
	Is the evaluation a legal requirement?	Yes/no.
	Evaluation a continuous or one-off activity?	Ad-hoc/continuous.
Context	Time (historical developments)	1-6
	Policy goals (intended outcomes)	0-4
	Policies in other sectors (interactions?)	0-4
	Unintended policy outcome(s)	0-4
	External events/circumstances	0-4
	Political environment/structures	0-4
	Geography	0-4
	Scientific findings (e.g., climate science)	0-4
	Evaluation methods	Record methods.
	Number of methods	Record number.
	Evaluation criteria	Record criteria.
	Number of criteria	Record number.
	Side effects	Yes/no
	Evaluation method tailoring	0-2
	Reflexivity	1-3
Interaction	For informal evaluation: Does informal evaluation attempt to identify and fill gaps left by 'formal' evaluation activities?	No gaps, gaps identified, gaps identified & addressed.
	Reference to evaluation studies conducted in other centres (but focusing on the same centre)	0-3
	Reference to (evaluation studies of) experiences in other centres.	0-3
	To what extent do 'formal' evaluations draw on information from 'informal' evaluations and vice versa?	0-3
	Is there a common metric (e.g., quantification?) that can be used to compare across governance centres?	0-4
	Are there key lessons/recommendations for others or the policy itself?	0-2
	If there are recommendations, is it clear whether/how the context matters?	0-2
	Ease of use	Executive summary (yes/no).
	If executive summary: (Categorical)	Hierarchy of information (yes/no) .
	Linguistic access	Summary in other language (yes/no).
	Availability of the evaluation	0-3
Not used/no data.	Evaluand – substance or process? (Categorical)	Policy substance, policy process, both policy substance & process.
	Evaluation budget	Budget.
	General comments	

4.4.6 Selecting a sample of evaluations

The large size of the 'population' of climate policy evaluations collected in the aforementioned database (comprising 618 evaluations) required selecting a sub-selection of evaluations for in-depth analysis with the coding scheme. Preliminary coding had revealed that applying the coding scheme in a systematic way often required reading large passages, if not the entire evaluation, in detail in order to extract the relevant information. Although the length of the evaluations varied, many had dozens, if not hundreds of pages of text. It was not uncommon to take several hours or even a full working day in order to code a single evaluation. In order to generate a sample, I used all informal evaluations and a random sample from the formal evaluations.

A crucial first step at this stage involved distinguishing between formal and informal evaluations for analytical purposes. The database overview above revealed that there are in principle two ways in which one could distinguish between formal and informal evaluations: focusing on evaluation funders or on evaluators themselves. Theoretically, the focus on the formal/informal distinction in this thesis derives from the Ostroms' ideas about self-governance, which is one of her foundational ideas (see Chapter 2). As Ostrom (2005) highlights, the crucial question on self-governance in monitoring and evaluation relates to the extent to which organizations can muster the necessary (i.e. principally the financial) resources in order to conduct rather expensive studies or, here, climate policy evaluations. With a view to self-governance, the key characteristic to distinguish between formal and informal evaluation is thus related to who pays for policy evaluations – information that can also in many cases be readily found in the evaluations, although none of the evaluations analysed here indicated any specific amounts spent on evaluation (the item was included in the coding scheme, but due to lack of data, no further analysis is possible).

Taking these aspects into account, I extracted 168 evaluations for in-depth analysis with the coding scheme, which the following section explains. Fritsch et al. (2013) had used an analogous method to draw a sample for coding from a population of impact assessment studies. In fact, given that there were only 84 informal evaluations, I analysed all of them, and drew a random sample of 84 evaluations

from the formal ones. I generated the sample of the formal evaluations with a random number generator in Microsoft Excel (asking the program to generate 84 random numbers out of a pool of 458). Alternatives to random sampling may have included random stratified sampling, which would have involved holding some attributes of the sample constant (such as, for example, the distribution of evaluations form the EU level, from Germany and the UK). However, given that earlier literatures did not suggest any a-priori factors that may have impacted significantly on the findings, I chose a standard random sampling approach, which in turn allows statistical extrapolation below (other sampling approaches may have made this more challenging).

4.5 Ethics

In the context of this thesis, ethical considerations relate to the fact that this research contributes to and is indeed part of wider political processes. Given that this study involved desk research of publically available evaluations, a whole gamut of ethical issues related to research with individuals did not emerge (see Frankfort-Nachmias & Nachmias, 1996). However, there were other important ethical issues to consider. The first is that many climate policies emerge from difficult political compromises, and their evaluation may thus touch on these sensitive political elements. In other words, to the extent that this thesis can also be understood as an intervention in debates about options for climate governance, policy evaluation, as well as the allocation of scarce resources to governance activities, it has the possibility to influence future outcomes. In order to address this issue, I endeavoured to work as diligently as possible and to discuss theoretical and methodological choices openly in Chapters 1-4, as well as in Chapters 8 and 9.

Furthermore, I collected and analysed a large number of evaluations from the public domain from a range of organizations with various stakes in climate governance. The key concern here was to generate findings and insights as accurately as possible, and to represent all organizations I could identify that had produced climate policy evaluations. In order to ensure fair and equal treatment of each study, I took great care to analyse each evaluation with the same standards and criteria, as

explained above. The goal was not to single out any one organization or evaluation, but rather to contribute to an understanding of broad trends and characteristics of climate policy evaluation practices in Germany, the UK and at EU level. Therefore, I only present aggregate data on evaluation characteristics in this thesis.

4.6 Reflections on the research process

This thesis has, in many ways, served as an intellectual groundswell of ideas and potential research directions that I have been able to address here, but also in flanking publications that accompany this thesis and consider important additional areas that were beyond the scope of these chapters, but nevertheless highly pertinent to the core questions addressed here and therefore also referenced throughout the thesis (see Jordan et al., 2015; Jordan et al., 2018; Schoenefeld & Jordan, 2017; Schoenefeld et al., 2018). Doing so in fact significantly enlarged the scope of this project, allowed me to receive early feedback and engage with the reactions of other scholars to my ideas. By the same token, it also meant that dealing with a significant larger workload than 'only' writing the thesis and keeping a reasonable balance between different tasks was challenging at times.

With a view to the specific analysis in this thesis, the heterogeneity of the climate policy evaluations in the database presented a challenge when creating a coding scheme that was both theoretically meaningful, and at the same time broad enough to incorporate many different approaches without running the risk of comparing 'apples and oranges.' The foundational ideas from polycentrism became a novel—and suitable—way to do so (see Chapter 2). In practical terms, this however also meant that it was difficult to gauge the time it would take to first assemble the database and then analyse a sub-section of it. With the knowledge I have now, I would have created a similar coding scheme, but endeavoured from the start to link it more closely with extant evaluation literatures and draw more on the work of others, an approach that I only developed relatively late in the thesis (and as I became more familiar with the evaluation literatures that Chapter 2 summarizes). Altogether, this process stretched well over three years and became therefore a highly ambitious undertaking with vast time requirements.

There are many key learning moments when embarking on a project of a length and duration that surpassed anything I had ever done before. During my research stay in Finland, Professor Mikael Hildén once remarked to me that 'getting a PhD is a lot about stamina,' and having gone through the experience, it is now very clear to me what he meant. The countless hours of database work and coding, as well as the process of writing up, tested my patience and willingness to engage with small, but often highly relevant or at least consequential, details. Then there is the process of 'zooming in' and 'zooming out,' or in other words, being able to see the thesis and its components as a whole, but also the nature and structure of the different elements. My supervisor Professor Andrew Jordan once cited his own PhD supervisor Professor Tim O'Riordan to say that writing a PhD (or a book for that matter) is 'like a symphony' and that creating harmony between the different parts is by no means trivial. In the process of creating my own symphony without too many dissonances, it has both been helpful to draw on relevant literature (e.g., Dunleavy, 2003), but also to use various 'navigation aids', such as drafts of the thesis abstract or the table of contents to keep all the relevant pieces well within sight.

Perhaps one of the most pertinent insights I am taking away from this process is that I have learned a great deal about creativity and especially how to stimulate it. This process is highly idiosyncratic and probably different for different people. While not all aspects of writing a PhD require creativity (some simply require stamina to get a task done), at some critical junctures, this is absolutely vital. I have experienced these points mainly towards the beginning of the PhD as I was working hard to delimit my research question and identify and understand the relevant literatures and various strands of argument. But the task of seeing beyond what is already there, recombining existing elements while creating new ones, is what ultimately requires high levels of creativity. It is a real art to identify a research question and approach that is ambitious enough to qualify as something novel, and yet still doable and realistic within the inevitable resource and time constraints. I learned how, on an individual level, creative moments are not equally distributed across a working day or working week and that the right mixture of activities, such as times of intense engagement with intellectual material, but then also times of relaxation and disengagement tends to generate better outcomes.

4.7 Limitations

As with any larger research endeavour, this project contains a number of limitations within which its analysis and findings have to be understood and contextualized. The first is that due to the novelty of many elements of the project – notably the development of polycentric governance theory, but also the climate policy evaluation database and the coding scheme – there were many areas where previous work as a reference point for my own research activities was severely limited. Wherever possible, I endeavoured to link with previous work (as for example with the coding scheme—see above), both theoretically and empirically, but I remain keenly aware that there are many areas where what I propose here could be further tested and explored. I sincerely hope that my work will serve as an impetus for future researchers to engage in relevant aspects of this research.

A second set of limitations emerged from the inevitable resource and time constraints that come with conducting a research project with a scope that has in other instances taken entire research teams to address (see for example Haug et al., 2010; Huitema et al., 2011). Working with limited resources required prioritizing various approaches and making strategic choices while ensuring continuously high research standards throughout this project. These included, for example, analysing a sample of the formally-funded evaluations (rather than the entire database), and testing the inter-coder reliability on a smaller sub-set of evaluations (rather than having a second coder for the entire process). I however hope that my database will allow others to do this work in the future and to continue exploring the various aspects of climate policy evaluation in the EU and well beyond.

Third, the novelty of the thesis and of the entire field of studying climate policy evaluation means that this thesis is, by and large, more descriptive than explanatory in nature. However, doing more descriptive, empirically-driven analysis is precisely what scholars working on climate change in the polycentric governance tradition have long called for (Jordan et al., 2015; Jordan et al., 2018; E. Ostrom, 2010c). In fact, deeper knowledge of the structure of polycentric (climate) governance is a vital pre-condition for further, causal analysis that I hope this thesis will support in the future (see Chapter 9). Making the climate policy evaluation landscape intelligible, and suggesting a theoretically-driven role for policy evaluation in polycentric

arrangements not only allows testing additional causal mechanisms, but also provides fertile ground for comparing the mechanisms hypothesized and tested here against others which could fulfil similar roles (see also Donaldson & Lipsey, 2006). For example, while policy evaluation may be one approach to spread knowledge and experience in a polycentric system, it is by no means the only one. Another mechanisms may include for example epistemic communities (see P. M. Haas, 1992) – again, this thesis allows a starting point for testing and contrasting a range of other mechanisms that may support polycentric governance.

Fourth, it is important to recognize that the data presented in Chapters 4-6 are based on overt statements of evaluation requirements and intent; it is possible that a number of evaluations do not explicitly mention legal requirements, but that the evaluation in fact responds to a legal requirement. Similarly, the coding scheme can of course only capture elements that are overtly (i.e. textually) available in the evaluations. By its very nature, it cannot capture other (and potentially more covert, but nevertheless important) elements, such as the process through which the evaluations emerged, relationships between funders and evaluators, or power struggles between different evaluation actors. As discussed above and in Chapter 1, the perspective in this thesis is undeniably a partial one, and it cannot make definitive statements about the nature of the process that generated the evaluations under analysis here. But doing so is a vital area for future research (see Chapter 9).

4.8 Conclusion

Studying policy evaluation from a polycentric perspective involves engaging deeply with the ideas of Elinor and Vincent Ostrom in order to not only understand key theoretical lines of polycentrism, but also their research philosophy. This thesis draws on these relevant debates and shows that the Ostroms advocated a combination of methodological pluralism and pragmatism, with which they aim to move beyond what are in their view entrenched methodological debates in political science. In their work, they actively combine normative and positive elements and advocate a dialectic relationship between them. This thesis engages seriously with their call for empirical work through collating a novel database of climate policy evaluations, and

subjecting the evaluations to analysis with a coding scheme that builds on the three foundational ideas on polycentric governance (see Chapter 2). Doing so involves a range of key methodological decisions, which this Chapter discusses (see also Chapter 9, which returns to some of these choices in light of the findings). The following three chapters discuss the detailed results from coding formal and informal evaluation.

Chapter 5 Formal Evaluation

5.1 Introduction

This chapter presents the results of in-depth coding of 84 evaluations funded by 'formal' actors; that is courts and scrutiny bodies, parliaments, governmental organizations, banks, or agencies, independent advice committees, research councils, universities/research institutes, and government (policy-makers). The data resulted from the analysis of the formal evaluations with the coding scheme that was based on polycentric governance theory (Chapter 2) and which Chapter 4 explains in detail. The 84 evaluations constitute a random sample from the 458 formal evaluations in the overall database. This chapter presents the results relating to the foundational ideas of self-organization, context, and interacting governance centres. Where appropriate, this chapter presents the data on the relative contribution of evaluations from the three governance centres considered in this thesis, namely the EU level, Germany, and the United Kingdom.

5.2 Self-organization

None of the evaluations that this chapter analyses are self-organized because formal (state) actors (or actors that receive state money) funded all of them. But the 'formal' category is by no means monolithic and contains considerable underlying variation, a core concept in polycentric governance (where heterogeneity is a key concern, see Chapter 2) and also important knowledge in order to understand the outcomes of formal evaluation below. Therefore, it is illuminating to unpick the 'formal' category into its component parts. Figure 5.1 unpacks the 'formal funder category' by funder location. The bar chart reveals that most formal evaluations (42.86%) are supported by funders from Germany, followed by the EU level (i.e. the

main EU institutions – see Chapter 4) with 40.48% and, to a considerably lesser extent, the UK with 16.67%.[29]

Figure 5.1: Funders by location

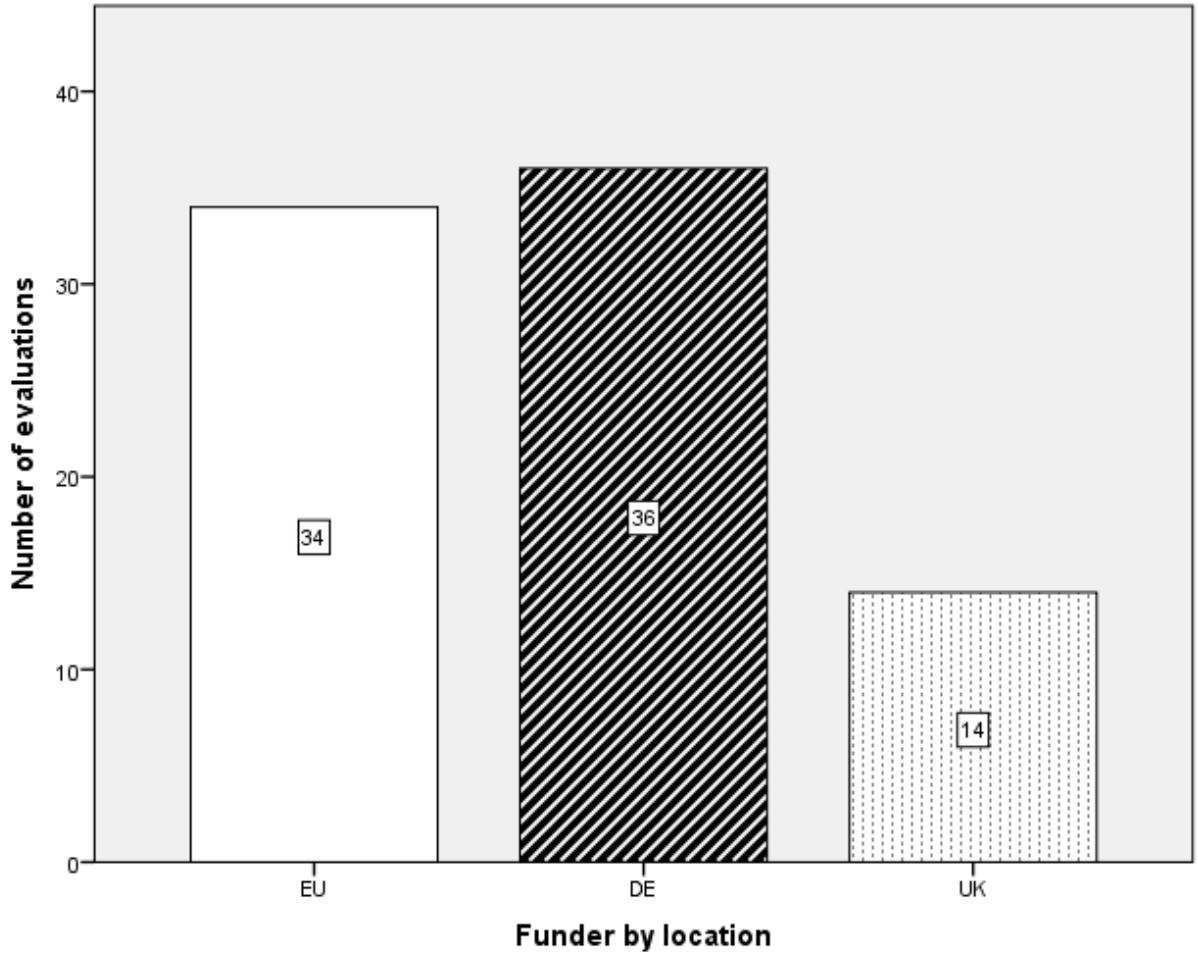
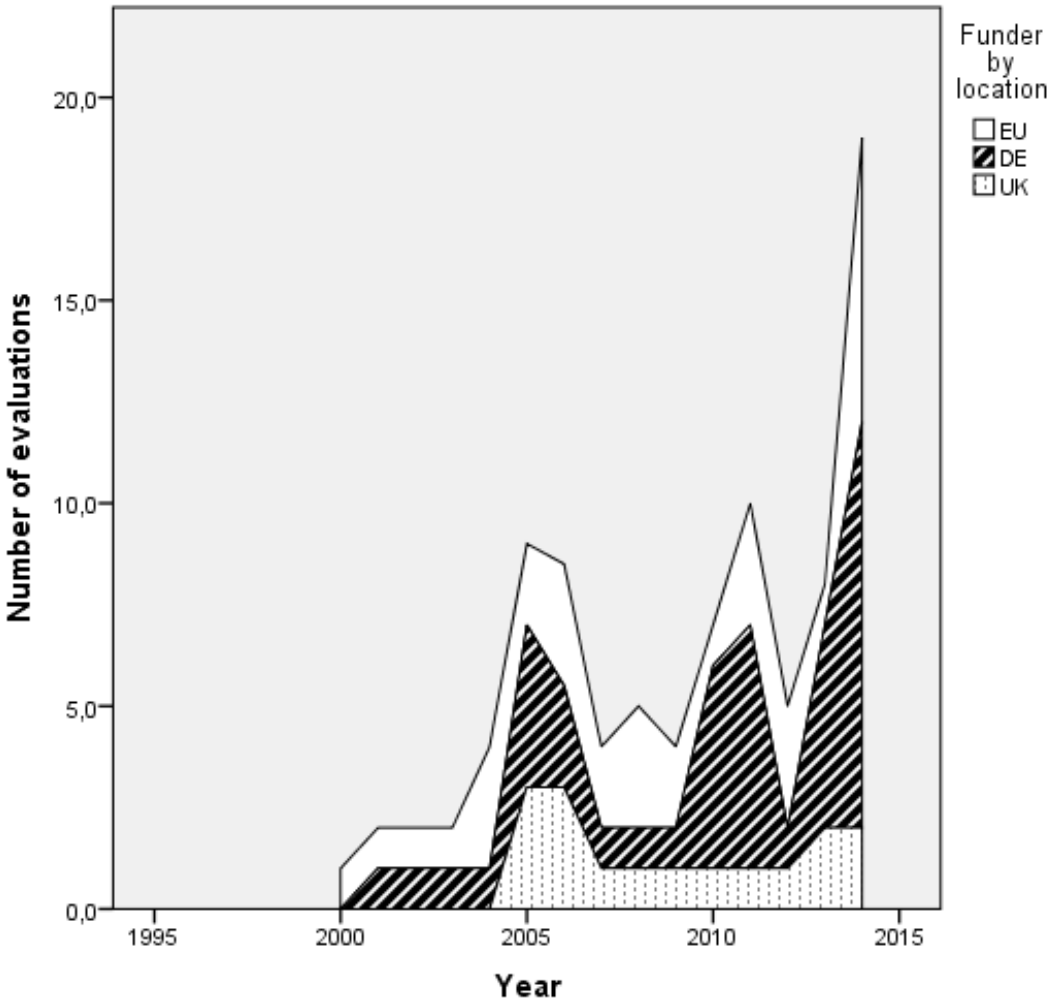


Figure 5.2, then, presents these country-based data over time. It shows that, first, formal actors commenced funding climate policy evaluations in the year 2000 (in the current sample – the total variation in the overall database is larger, see Chapter 4). While overall the number of evaluations is growing in line with the overall database (Chapter 4), there are also considerable variations in the number of formal evaluations at EU level, in Germany, and in the UK over time. Figure 5.2 shows that evaluation peaked around 2005, 2011, and around 2014 (although given

[29] The percentages add up to slightly more than 100% because of rounding errors.

that the data here only extend to 2014, the latter point must be understood much more tentatively). In most cases, EU level funders appear to provide more continuous funding, and assure that total evaluation output never goes much below five climate policy evaluations per year after 2005. By the same token, evaluations funded by actors in Germany wax and wane considerably in number over the past decade or so. The number of evaluations funded by actors in the UK remain at a comparatively low but stable level, even though they commence later than those funded by actors at the EU level or in Germany.

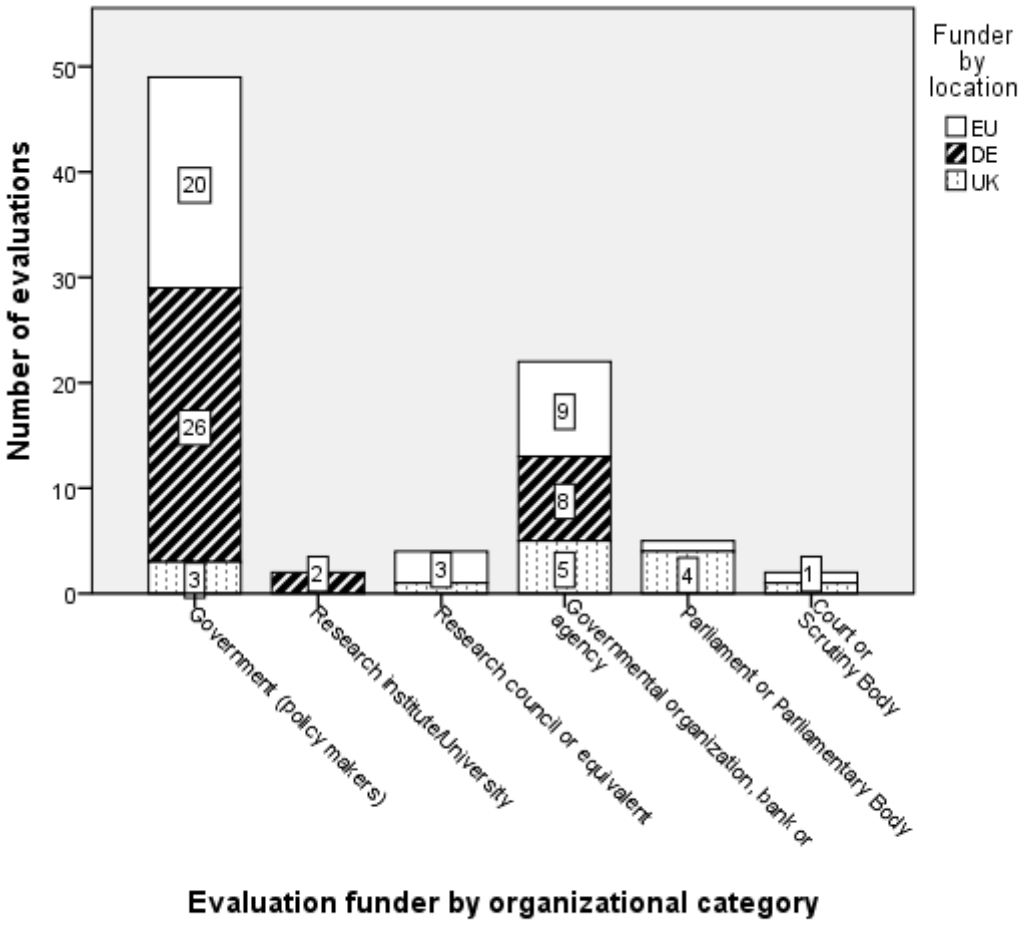
Figure 5.2: Evaluations by year and funder location



Then there is the question of what types of formal organizations actually fund the evaluations. Looking at the overall height of the bars in Figure 5.3 reveals that governments (policy-makers) fund the overwhelming majority of evaluations,

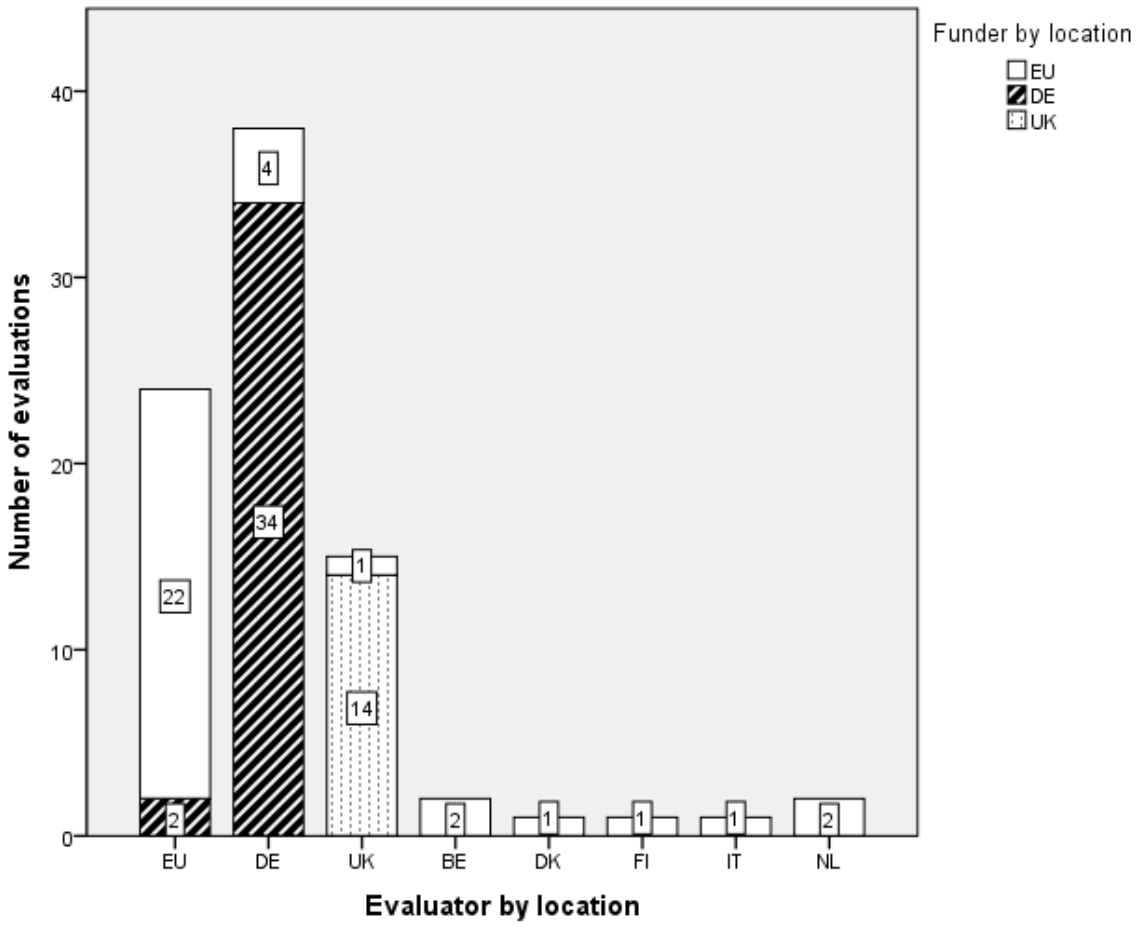
followed by governmental organizations (such as agencies or public banks), and still a smaller share by parliaments, research councils, and courts, as well as research institutes or universities. This set includes courts, which may be independent from government, but are still part of the state, and therefore bound to legal requirements (such as previously defined policy targets), even though they of course also have a role in interpreting the law. Given that the number of climate policy evaluations conducted by courts or public scrutiny bodies is very low, any potential distortions are likely to be of limited effect. Figure 5.3 thus also points to the insight that the internal differentiation of the 'formal' category matters. Note that German funders only include governments, research institutes/universities and governmental organizations/banks, while the range of funders at the EU level and in the UK is broader and includes parliaments, courts and research councils.

Figure 5.3: Evaluation funders by organizational category



Turning from evaluation funders to those who actually conducted the evaluations (i.e. the evaluators), Figure 5.4 presents the number of evaluations by the location of the evaluators who produced them, as well as by the location of the funders. The first thing to note is that evaluators located in Germany produced thirty-eight formal evaluations (the total number of evaluations contained in the second bar), which is the biggest sub-group in this sample. Evaluators at the EU level (recall that this category includes cases where evaluators from several European countries teamed up to produce the evaluation—Chapter 4) were the second-largest with 24 evaluations altogether. Evaluators located in the UK produced 15 evaluations, which is significantly less than those from the EU level, and less than half of the evaluations produced by evaluators in Germany. Evaluators from other countries, including Denmark, the Netherlands, Belgium, Italy and Finland, also conducted a handful of formal evaluations.

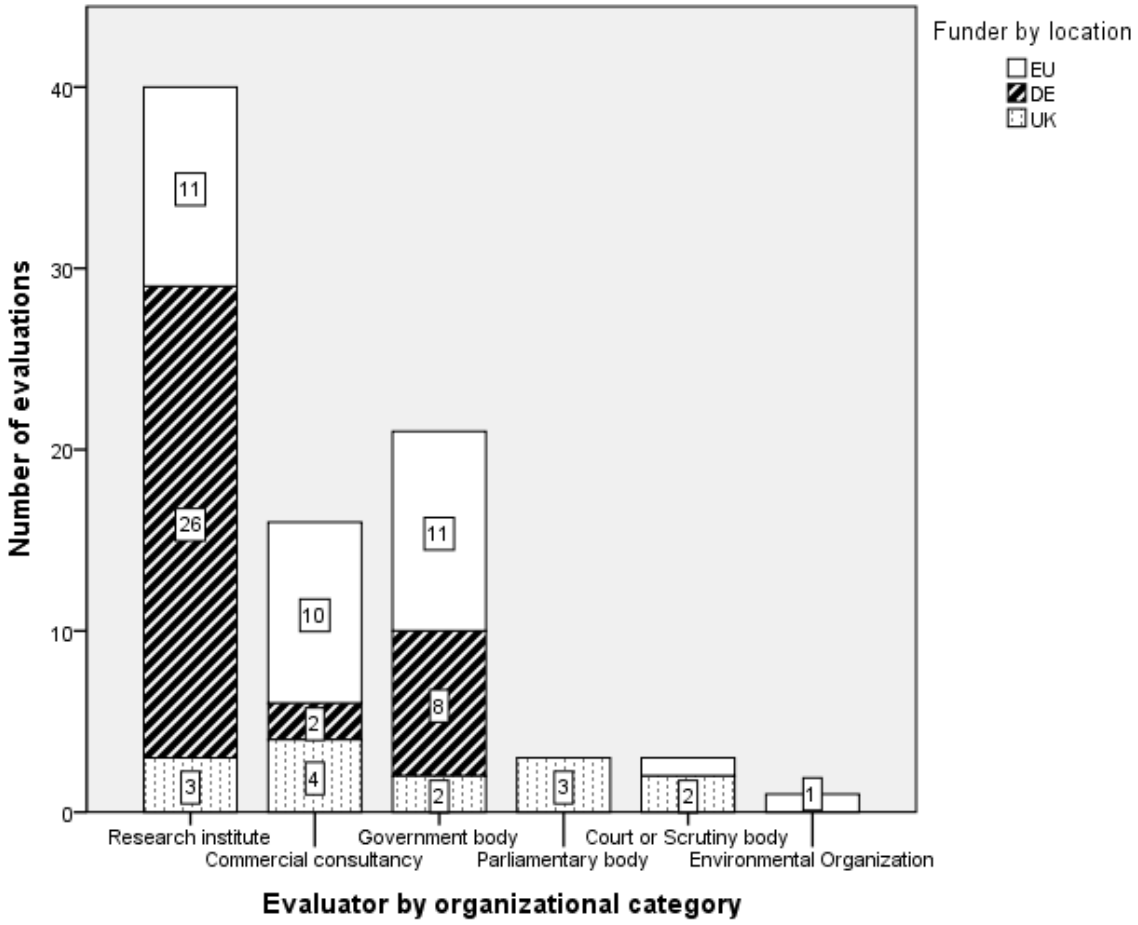
Figure 5.4: Evaluations by evaluator and funder location



The second thing to note is how the combination of the location of the evaluation funder and the evaluator plays out—which in the context of Figure 5.4 involves considering the differently shaded elements of each bar. By and large, Figure 5.4 suggests that evaluation funders tend to fund evaluators within their own governance centre, so that EU level actors funded an overwhelming majority of evaluators at the EU level (only two funded by German actors and none by UK actors), German funders funded 34 out of 38 evaluations conducted by evaluators in Germany (the remaining four were funded by EU actors) and, even more extreme, actors in the UK funded 14 out of 15 evaluations conducted by UK based evaluators. These findings suggest that funding and conducting formal climate policy evaluations remains, by and large, a national or EU level affair with relatively little cross-border interaction via funding evaluators from other governance centres in this sample. But it also points to an important role of the EU, which funded evaluators in other governance centres to a limited extent, such as Belgium, Denmark, Finland, Italy, and the Netherlands.

In addition to the location of the evaluators, Figure 5.5 (below) reveals the type of organizations that that formal actors funded in order to conduct their evaluations (recall that self-funding is possible). It demonstrates that formal actors mainly fund research institutes to conduct evaluations, followed by government bodies and commercial consultancies. Parliamentary bodies, courts, and environmental organizations rarely receive funding in order to conduct evaluations. Then there are differences in who evaluation funders choose: whereas EU evaluation funders used research institutes, commercial consultancies and government bodies in about equal numbers (with negligible funding for other types of evaluators), German actors make stronger use of research institutes, followed by government bodies and a very small number of evaluations conducted by commercial consultancies. Funders in the UK spread relatively evenly across the categories, but did not fund environmental organizations to conduct evaluations.

Figure 5.5: Evaluators by organizational category



A slightly different way of looking at the current sample of formal evaluations is to consider the governance centre in which the evaluated policy is located (rather than the location of the evaluators in the previous chart). Figure 5.6 presents the respective data. Looking at the height of the three bars, it suggests that the greatest number of climate policies under evaluation were located at the EU level (36 in total), followed closely by climate policy in Germany (34) and, to a considerably lesser extent, the UK (14). Combining these insights with the location of the funder (i.e. considering the different patterns of each bar) shows that, again, formal evaluation funders mainly funded evaluations focusing on their own governance centre in terms of their content. Formal German evaluation funders only funded evaluations of climate policies in Germany. A similar trend is evident for the EU and the UK levels. Considering Figure 5.4 and Figure 5.6 together reveals that, for example, German funders mainly funded evaluators in Germany focusing on German policies. A similar pattern applies to the EU level and to the UK.

Figure 5.6: Evaluations by location of the policy under evaluation

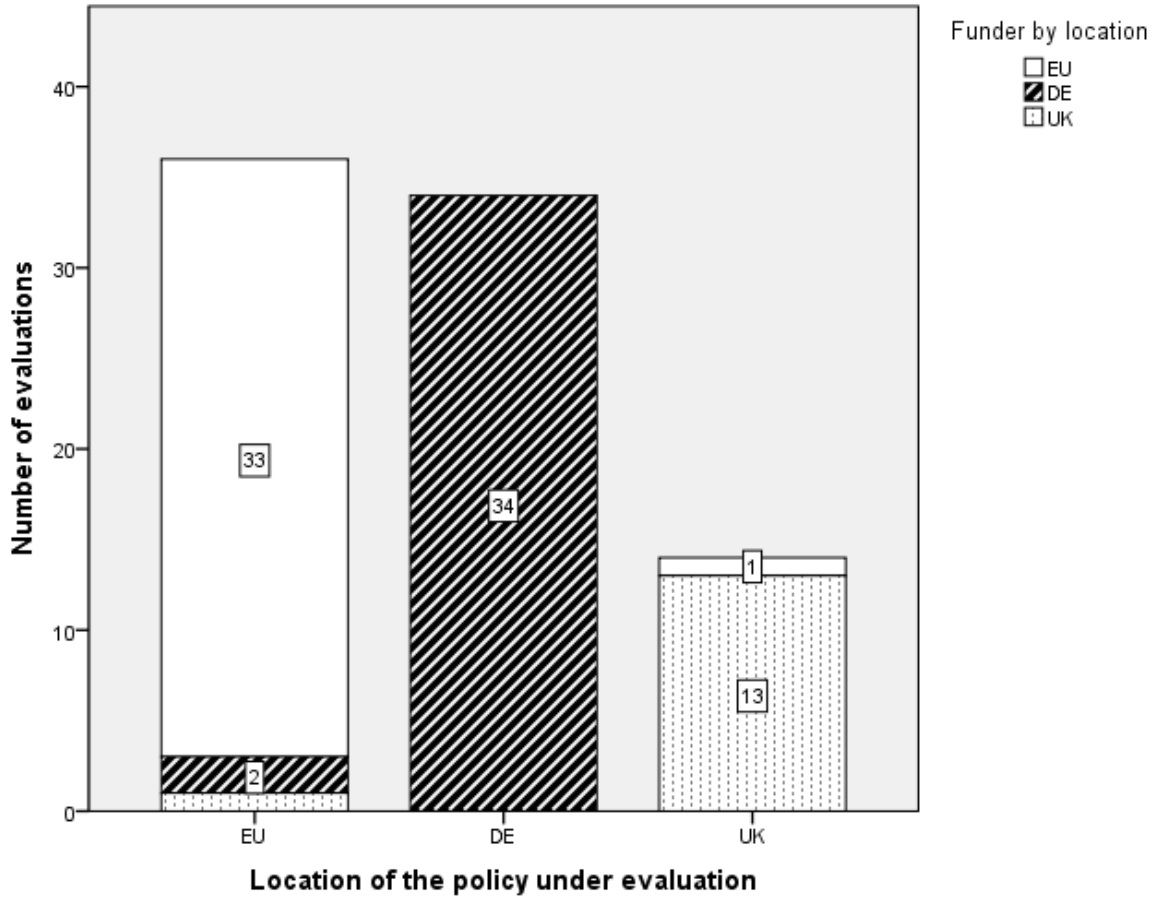
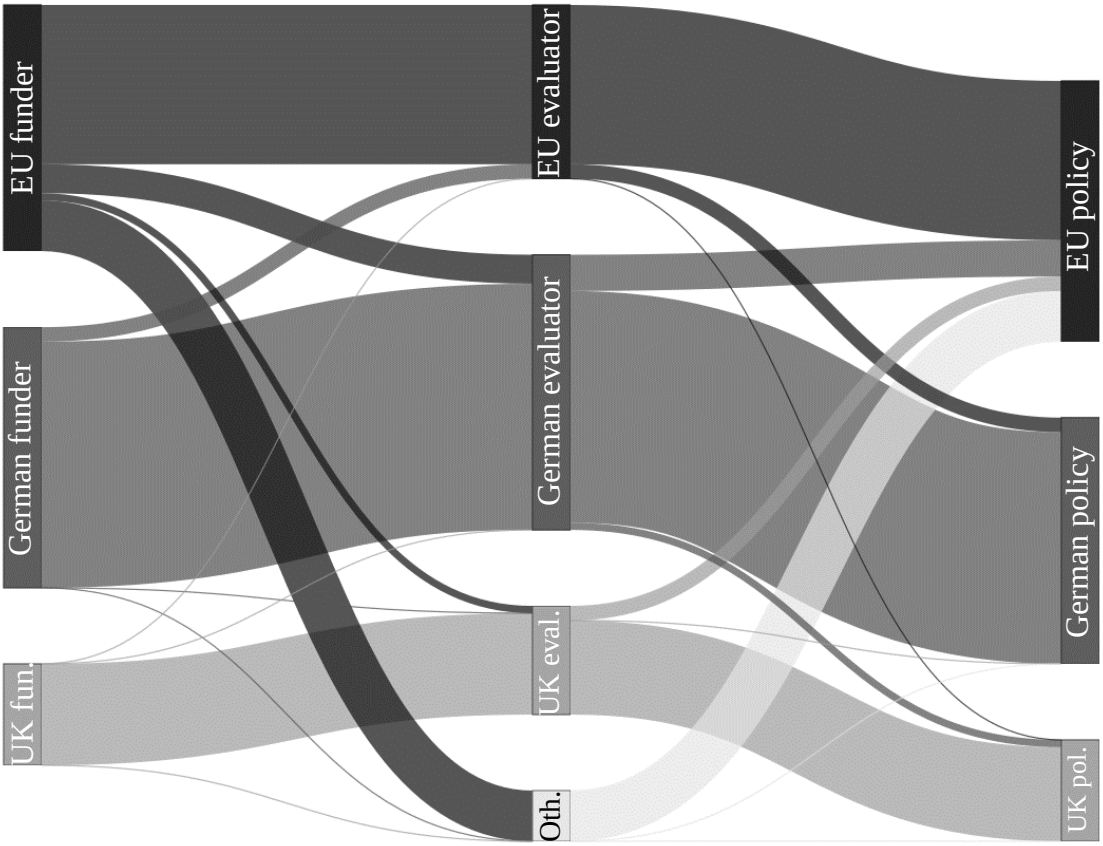


Figure 5.7 combines the data on the location of the funders, the evaluators, as well as the policy under evaluation. Analogous to Figure 4.7 in Chapter 4, the diagram contains the three perspectives on evaluation, and the thickness of the connectors between them represents the number of evaluations with the respective characteristics. A readily visible feature is that funders in the UK mainly funded evaluators in the UK in order to evaluate UK climate policy, with similar trends for the EU level, as well as Germany.

Figure 5.7: Location of funders, evaluators and policy under evaluation

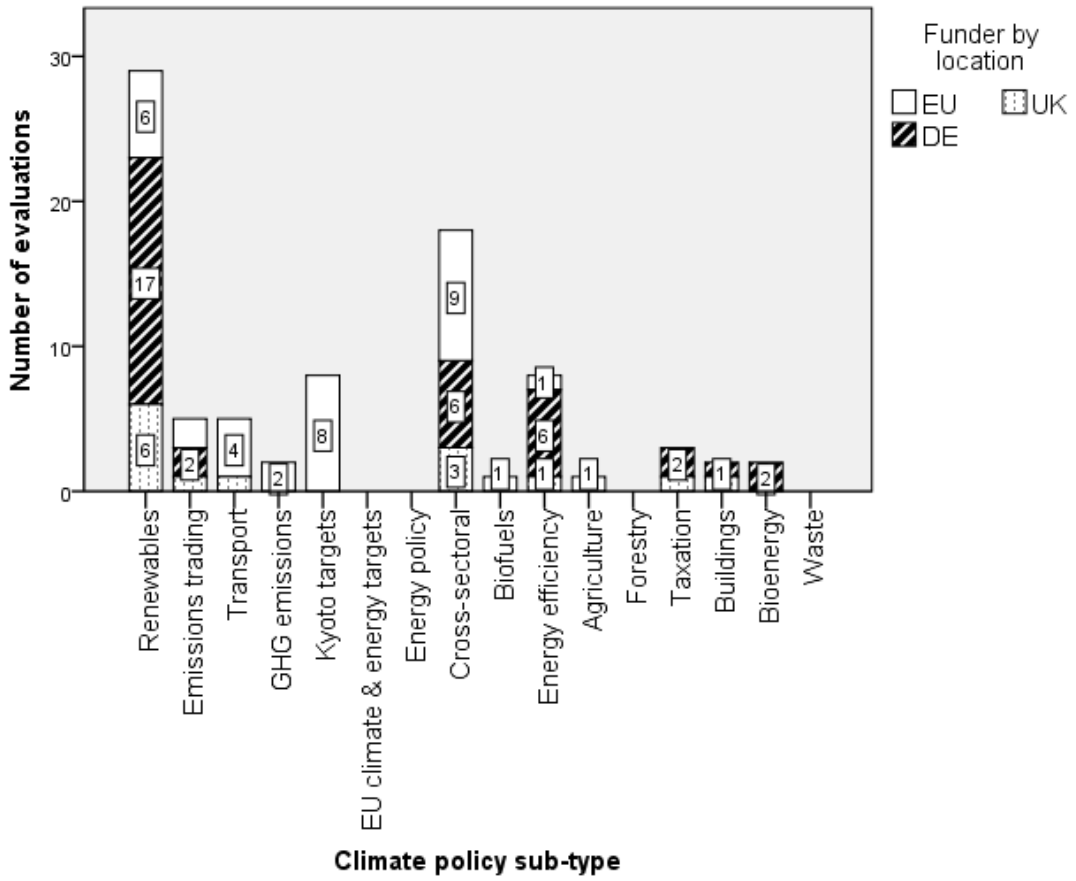


Note: the thickness of the bars represents the number of evaluations with the respective characteristics.

Finally, Figure 5.8 groups the evaluations by climate policy sub-type. Similar to the overall database (Chapter 4), the main substantial focus of formal evaluations is renewables policy, followed by cross-sectoral analyses (i.e. focusing on more than one of the individual climate sub-policies listed in Figure 5.4), and energy efficiency. Notably, evaluation against policy targets tends to focus on those emerging from the Kyoto Protocol, rather than the EU-specific targets (even though the two are of course closely related). There is also a marked paucity of evaluations in sectors that have considerable greenhouse gas emissions, such as transport, agriculture, or buildings. Formal climate policy evaluation thus mainly focuses on a few climate policy-sub types, namely renewables, cross-sectoral issues, energy efficiency and the Kyoto targets and does not consider some sectors at all, such as waste or forestry. Looking at the evaluation funders again shows that EU funders exclusively funded

evaluations on the Kyoto targets, and German funders focusing on renewables, cross sectoral, and especially energy efficiency policies.

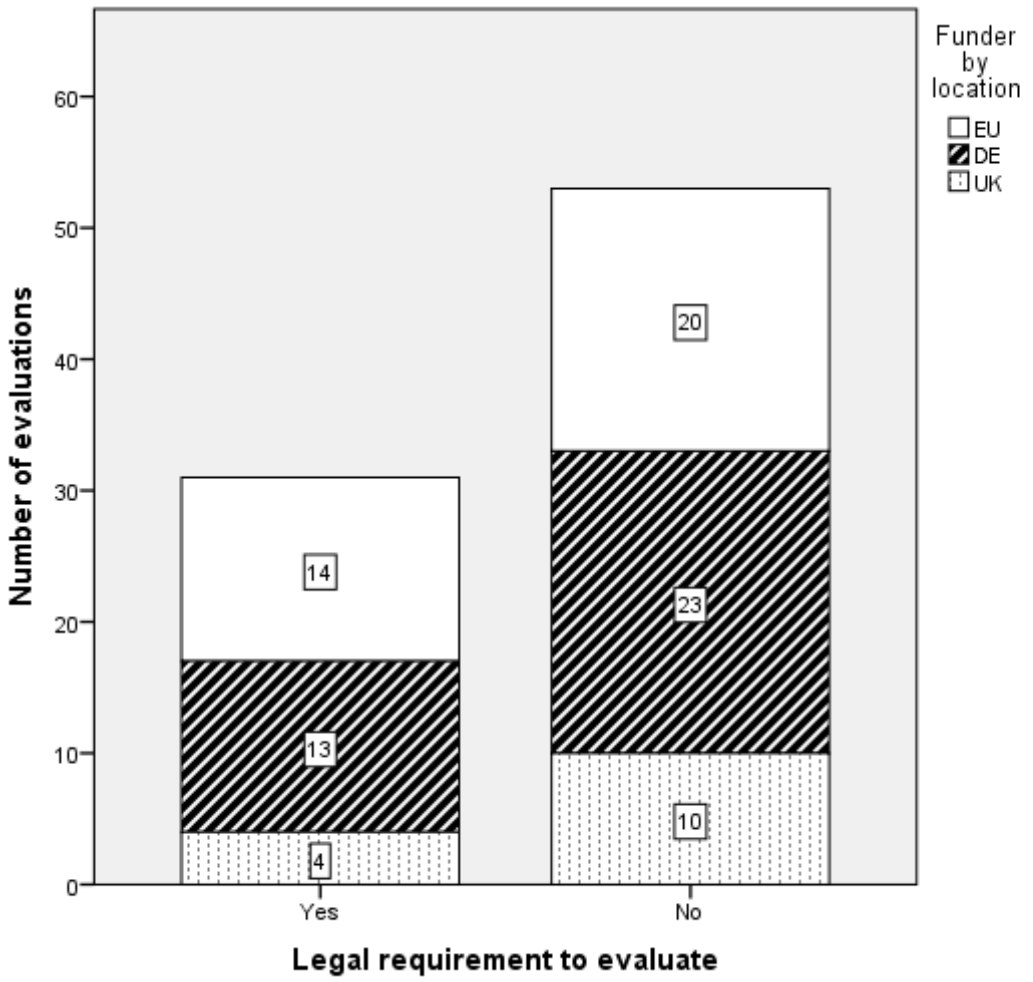
Figure 5.8: Evaluations by climate policy sub-type



But what stimulates formal actors to fund evaluations? Figure 5.9 shows to what extent formal actors were following legal funding requirements in their climate policy evaluation activities. As Chapter 2 discussed, the presence or absence of legal requirements may also be an indicator of how 'spontaneous' or 'self-organizing' the evaluations were, even among state-funded ones. Figure 5.9 reveals that a clear majority of the formal evaluations (63.10%) did not respond to a legal requirement, but were conducted for other reasons, such as a desire to improve policies through learning (for a fuller discussion of evaluation purposes, see Figure 5.23 below). By the same token, this means that 36.90% of the formal evaluations did respond to a legal requirement to evaluate. Recall that by and large, this item was coded by detecting whether or not there was any indication of a legal requirement in the

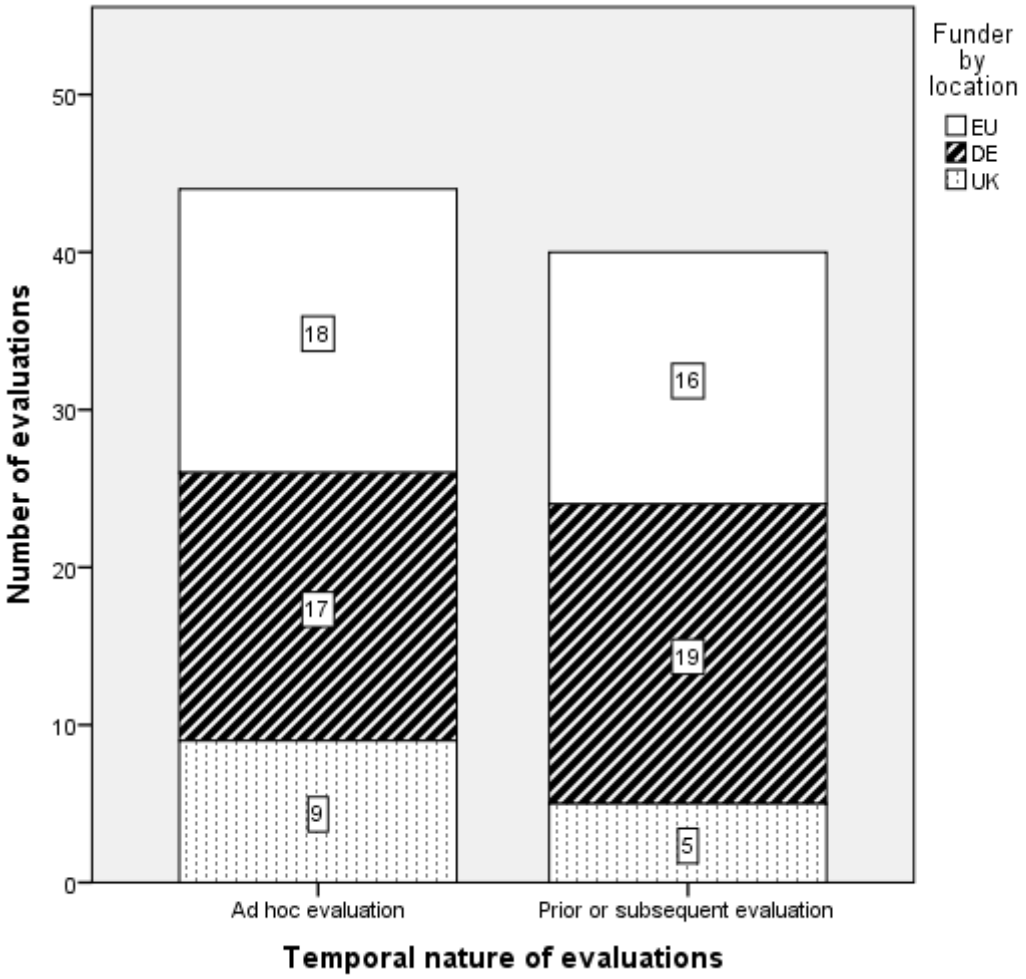
evaluation (such as reference to an evaluation clause in a legal document). The distribution by the location of the evaluation funder is about proportional.

Figure 5.9: Evaluation responding to a legal requirement



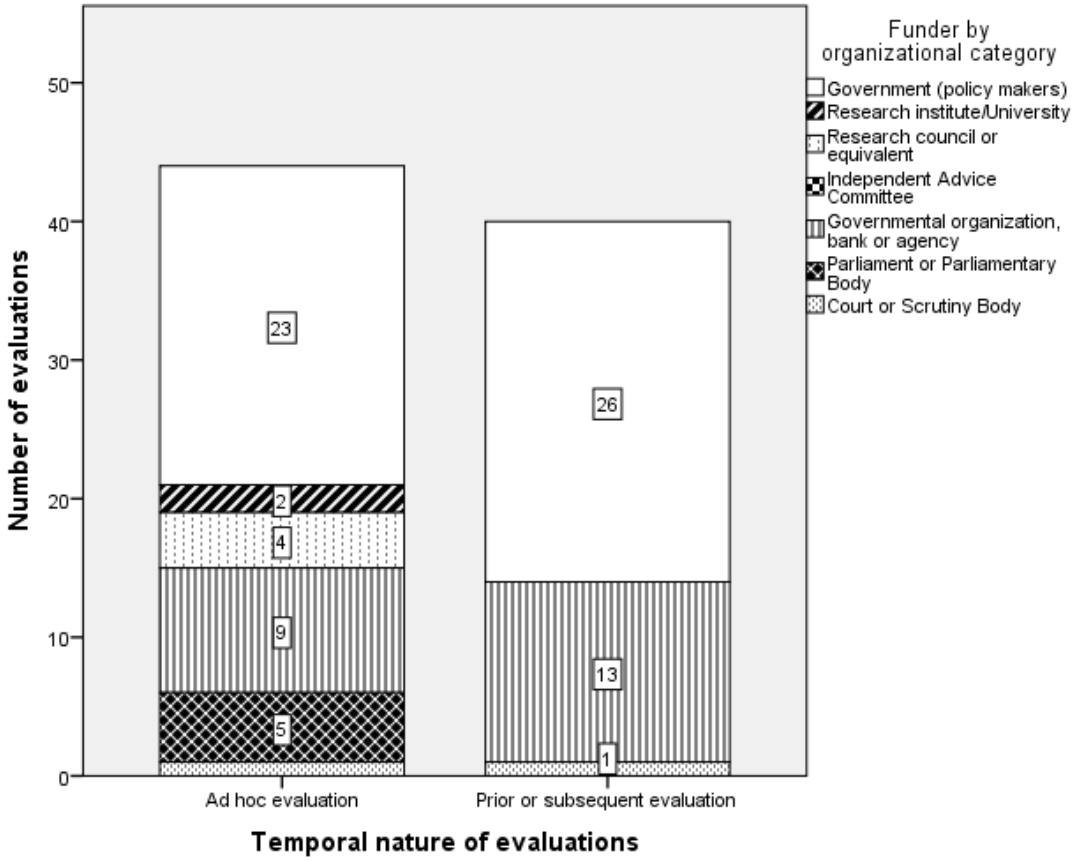
But what is the temporal nature of evaluations that formal actors funded? Are the evaluations part of larger and continuous evaluation exercises or are they rather ad-hoc? Recall that in dynamic, polycentric environments, continuous evaluation is thought to be particularly important (see Chapters 2 and 8). Figure 5.10 shows that there is by and large a balance between formal ad-hoc evaluations and those that link with earlier or later evaluations, for example as part of an ongoing series. In other words, formal funders funded both a significant number of ad-hoc, but also continuous evaluations in this sample. The distribution by the location of the evaluation funder is broadly proportional.

Figure 5.10: Temporal nature of evaluations by funder location



Given that the formal evaluation funder category is internally differentiated in terms of the organizations that fund evaluations (see above), Figure 5.11 considers the same data by the organisational category of the evaluation funders. It shows that continuous climate policy evaluation in this sample is in practice only funded by three types of funders: governments, governmental agencies or banks, and (to a considerably lesser extent) courts. Notably, parliaments, independent advice committees, research councils, and research institutes/universities are less involved in evaluating climate policy continuously over time.

Figure 5.11: Temporal nature of evaluations by funder category



While of course all of the evaluations analysed here were, by definition, *not* self-organized because formal (state) actors funded them (see Chapter 4 and above), this section has revealed that it is worth further unpacking the underlying characteristics of the formal category – including by the location of the evaluation funder – in order to work towards a deeper understanding of its internal structure.

- Crucially, formal evaluations tend to be very much anchored in one location in terms of who funds them, who conducts them, and what they focus on in substantial terms. In other words, EU level funders tend to, by and large, fund evaluators at the EU level who focus on EU level policies. Overall, there is a clear tendency for formal actors at the EU level and in Germany to lead on financing climate policy evaluation compared to the UK.

- There is a strong focus on renewables and targets, but formal evaluations do not cover all climate sub-policies equally. By and large, the evaluations presented in this chapter follow the overall distribution of the complete database (Chapter 4).
- Formal evaluation does not always depend on legal requirements. Rather, a majority of climate policy evaluations have been conducted for motivations other than mere legal requirements (see below).
- Formal evaluation funders have funded both ad-hoc evaluations and continual exercises in about equal numbers. The latter are particularly important when seeking to understand and track climate policy developments over time.

5.3 Context

This section presents analyses of variables that mainly relate to contextual elements. Figure 5.12 lays out the results from eight separate context-related variables used in the coding scheme (see Chapter 4). The bar chart in the top left corner of Figure 5.12 (Chart A) focuses on the length of time considered in the evaluation, which may include aspects of policy history or longitudinal data contained in the evaluation (see Chapter 4). It reveals a relatively broad spectrum of attention to time in the formal evaluations. There are relatively few 'snapshot evaluations' (9.52%; for a definition, see Chapter 4), meaning that most evaluations consider five or more years in the analysis or policy history. In fact, the biggest number of evaluations (highest bar) considers a time span of more than 20 years. Following the arguments in Chapter 2, considering a long time span would, in principle, be a good starting point in order to unpick potential ways in which a climate policy interacts with its context. Chart A in Figure 5.12 further reveals a relatively uniform and proportionate distribution of time horizons considered in evaluations funded by actors at the EU level and in Germany. The UK appears to mainly fund either evaluations with a very short or a very long time horizon.

Figure 5.12: Contextual variables in formal evaluations

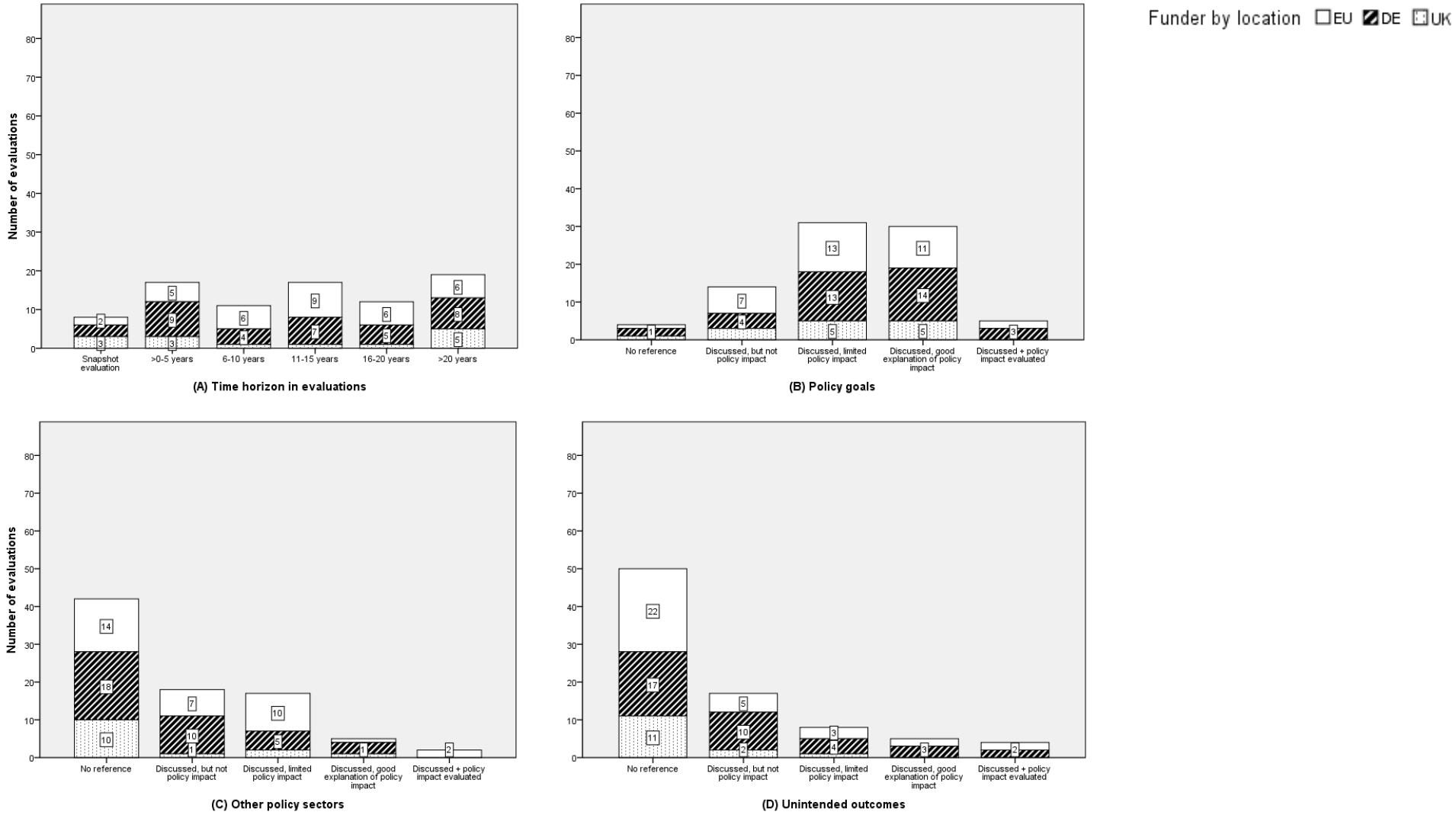


Figure 5.12 (continued)

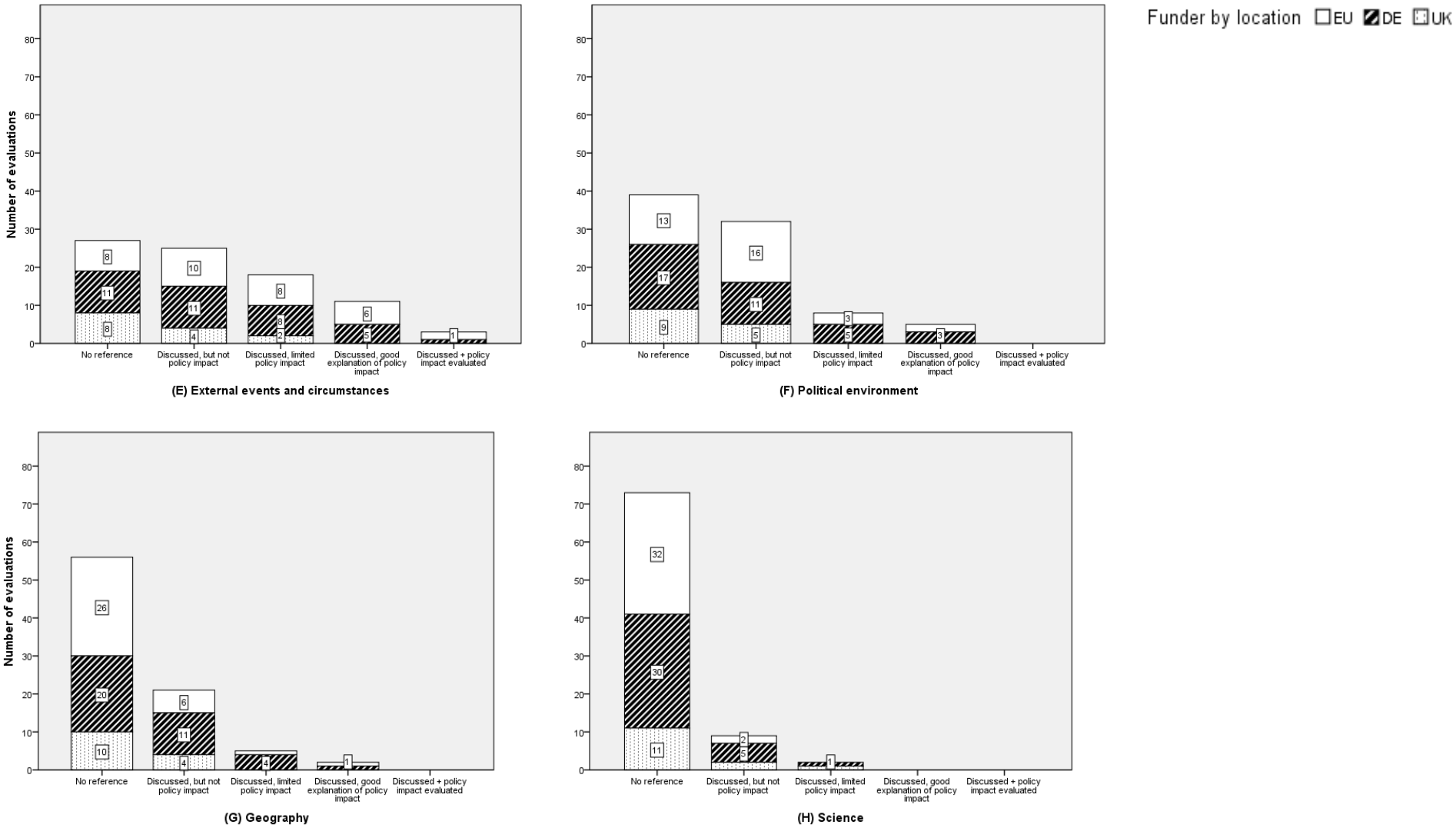


Chart B in Figure 5.12 and all the remaining ones in the figure assess attention to the contextual factors on a scale of 0 to 4.[30] More specifically, Chart B assesses the extent to which the evaluations considered policy goals (i.e. goals of the individual policy, but also more general targets to which the policy contributes, such as national greenhouse gas reductions). The bar chart shows that most formal evaluations (95.24%) made some reference to policy goals, but only 41.67% of the evaluations engaged with policy goals in any more depth and related the latter to the policy effects that the evaluation assessed. The number of evaluations that considered policy goals extensively (i.e. a score of 4 on the scale explained above) remains relatively small (five evaluations). The distribution of formal evaluation funders shows that EU level, as well as German and UK based actors funded evaluations in relatively equal proportions in all categories, with the exception that UK based funders did not support evaluations that scored very high on engaging with policy goals (see the shading in the last bar).

Chart C in Figure 5.12 demonstrates that about half of the evaluations made no reference whatsoever to policy interactions with other sectors—this includes both interactions across climate policy sub-policies (such as renewables and emissions trading), and interactions with other policy sectors (such as health policy—see Chapter 2 for a more detailed review of the core contextual variables), so pointing to generally very limited attention to contextual effects in relation to this particular dimension. A prominent example is the interaction between the EU Emissions Trading System and the Kyoto flexibility Mechanisms. Some evaluations considered interactions with policies in other sectors, but the number of evaluations that looked at the policy impact of linkages and interactions in greater detail remain just a handful in this sample. The distribution of funding remained relatively even, although there appears to be a slight tendency for EU level funders to support more

[30] Recall that this variable and all the remaining variables in Figure 5.12 were scored on a 0-4 scale, where 0 = no reference to dimension; 1 = dimension discussed, but no explanation of how this dimension impacts policy outcomes; 2 = dimension discussed, but limited explanation of how this dimension impacts policy outcomes; 3 = dimension discussed, and good explanation of how this dimension impacts policy outcomes; 4 = dimension discussed and impact on policy outcomes evaluated extensively (for further details, see Chapter 4 and Appendix 3).

evaluations that look across different sectors. From a polycentric perspective, this points to a key, overarching role for the EU level that differs from policy evaluation in the nation states of Germany and the UK (for a fuller discussion, see Chapters 8 and 9).

Chart D in Figure 5.12 presents the number of evaluations with different levels of attention to unintended policy outcomes.[31] The main message from this chart is that the vast majority of the formal evaluations make no or very few references to unintended policy outcomes. Formal evaluations that assess unintended policy outcomes in greater detail—that is, receiving a score of 3 or four on the scale explained above—remain far and few (10.71%). However, given that unintended effects may particularly become evident in other sectors or outside the focus area of the policy, this finding fits with the previous two bar charts, in that if there is little attention to effects in other sectors, the evaluations may then also be unlikely to detect significant unintended side effects occurring outside the focus area of the evaluation. With regards to the formal funders, it is noticeable that the few evaluations that do engage more deeply with unintended policy effects tend to be funded by actors at the EU level or in Germany. Notably, the UK did not produce a single formal evaluation that retuned 'good' (score 3) or 'extensive' (score 4) attention to unintended side effects (the two bars on the right).

External events and circumstances may also affect a climate policy and its outcomes in unforeseen ways. Therefore, Chart E in Figure 5.12 presents the extent to which formal evaluations engage with the impact of external events and circumstances on climate policy outcomes. A good example of an external event considered in the evaluations is the global recession, which started with the financial crisis in 2008, and then gradually morphed into a full-blown economic and sovereign debt crisis in the EU. By and large, Chart E reveals that attention to external events and circumstances is rather limited, and more detailed analyses remain rare. Only 13 evaluations engaged well (score = 3) or extensively (score = 4) with external events and circumstances and, notably and in line with our previous findings, none of these

[31] For definitions and operationalizations, see Chapter 4.

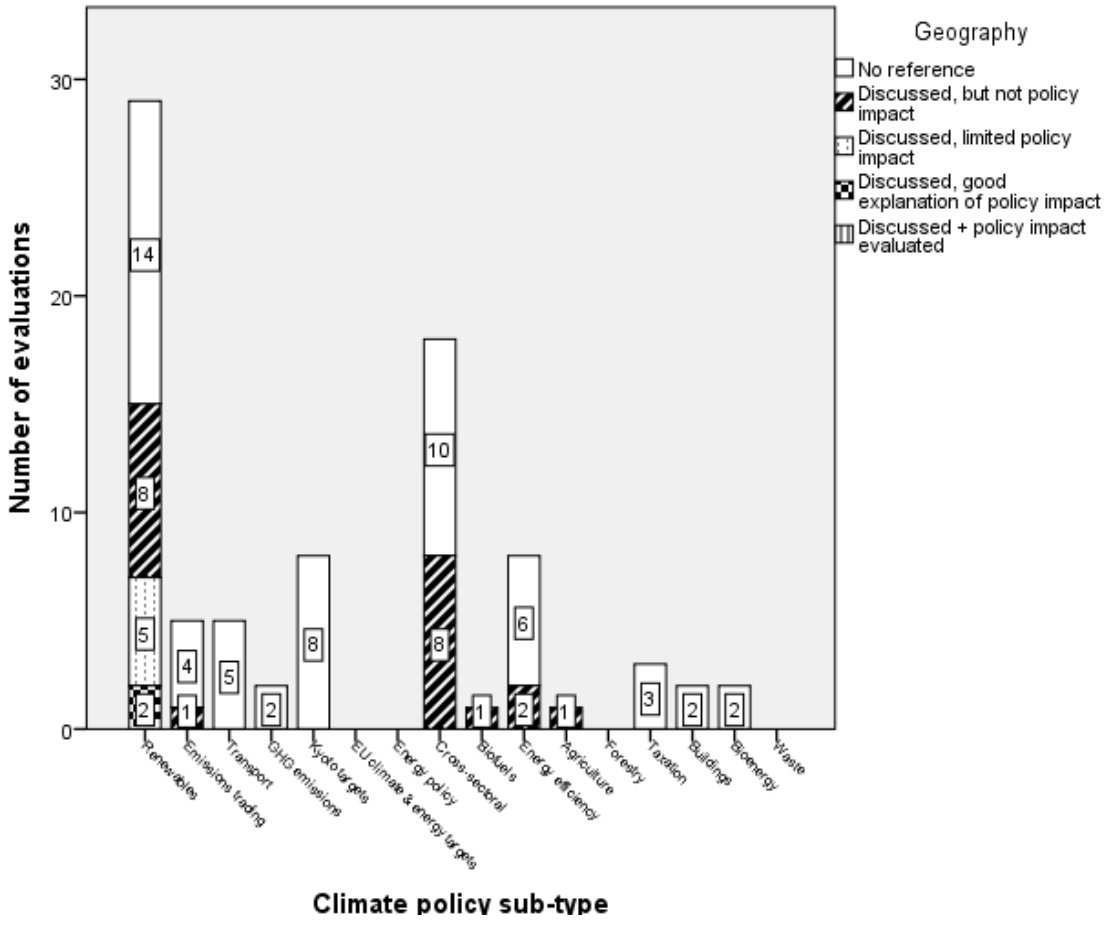
evaluations were funded by actors from the UK, but from actors at the EU level and in Germany.

Chart F presents data on the extent to which formal evaluations incorporated an assessment of the political environment in which any policy is placed (for the operationalization, see Chapter 4), including institutions and political shocks, such as elections or other significant events. A current example of such effects may be the election of Donald Trump as President of the United States of America and his subsequent decision to withdraw the USA from the Paris Agreement, a decision that may affect the dynamics of international climate policy-making. In this sample, the vast majority (84.52%) of the formal evaluations did not consider how the political environment affected climate policy outcomes and, notably, there are no evaluations that analysed the policy effects in great depth (i.e. a score of 4 on the relevant scale). But evaluations may be subject to different political influences or pressures, notably because the political systems in these three governance centres differ considerably (with the EU being a supra-national organization, Germany a federal state, and the UK a unitary state). Omitting the political context thus potentially leaves significant gaps in an evaluation. Reviewing the individual bars in Chart F in Figure 5.12 reveals that it is worth considering these data by governance centre: no formal actors in the UK have funded evaluations that analyse the impacts of the political environment on climate policy outcomes – the few studies that do so to some degree have exclusively been funded by actors from Germany and from the EU level, an overall pattern that fits the discussion of the charts in Figure 5.12.

Chart G in Figure 5.12 presents the next contextual variable considered in the evaluations, which is (mainly physical) geography, such as, for example, the availability of tidal energy as a function of the length and nature of a coast line that a governance centre exhibits. Many climate policies, such as support for renewables, depend to a significant extent on the availability of certain geographical conditions – it is, for example, not possible to use geo-thermal energy in all places at a reasonable cost. Chart G reveals the level of attention to geography in formally-funded climate policy evaluations on the already familiar scale. Across the formally-funded evaluations, the characteristics of the physical geography appear to be only of limited interest. Well over half of the climate policy evaluations did not pay any attention at all to geographical aspects, or mentioned them but did not discuss the policy impact.

However, it is important to recognize that in some cases, geography may play a more significant role, such as in the case of opportunities for generating renewable energy than in others, such vehicle emissions standards. Figure 5.13 (below) therefore splits up the data by climate policy sub-type in order to test the notion that geography may matter more for some policies than for others. The findings from this figure are relatively straightforward – it represents the climate policy sub-types on the x-axis and the number of evaluations on the y-axis. The different shading on the bars encompasses the points on the 0-4 scale used to code attention to geography. As discussed above, the only policy area where evaluators discussed geographical aspects in any meaningful depth beyond quickly mentioning it in passing is renewables policy, but even there, this only applies to seven evaluations.

Figure 5.13 Attention to geography in evaluations by climate policy sub-type

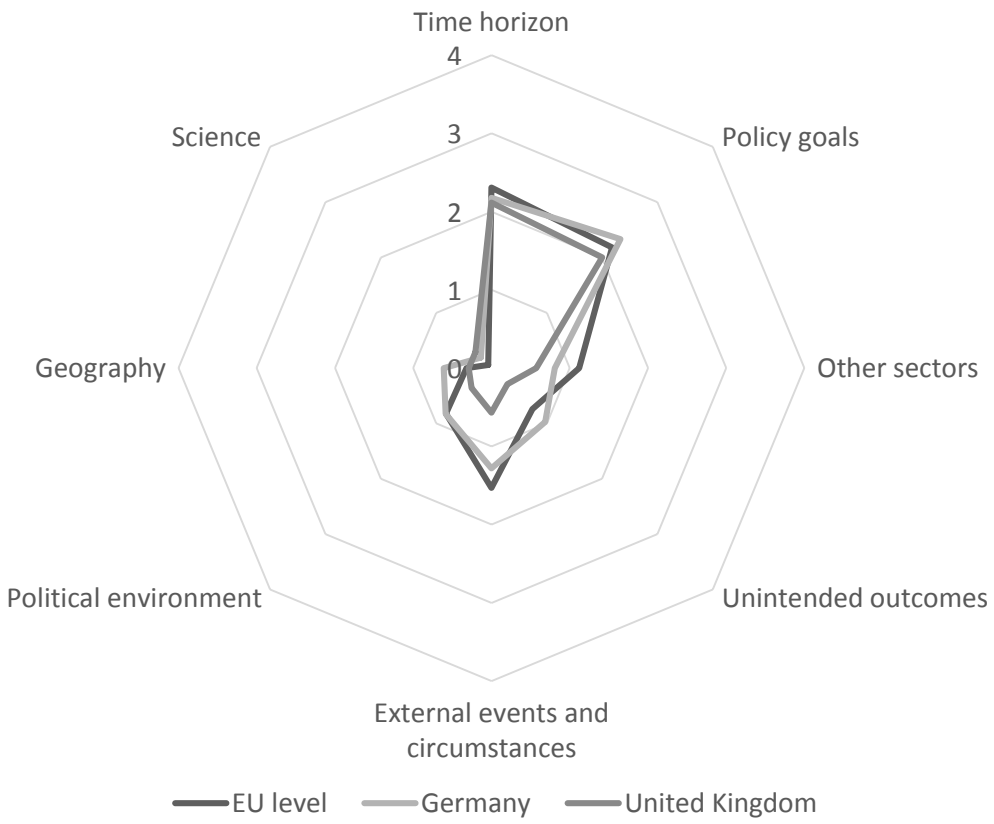


Returning to the discussion of Figure 5.12, the last contextual variable considered on the 0 to 4 scale (see above and Chapter 4) includes references to (climate) science and the scientific backdrop against which climate policies are made over time (Chart H). This includes, for example, references to the Intergovernmental Panel on Climate Change (IPCC) reports or what kinds of greenhouse gas emissions reductions are necessary in order to avoid the worst consequences of climate change. Chart H shows that the vast majority of climate policy evaluations do not reference the findings of climate science at all (86.90%), and the very few that do tend to do so in a fairly cursory way. There are therefore no formal climate policy evaluations in this sample that engage with the scientific backdrop in any significant detail (i.e. scores of 3 or 4 on the scale described above). The distribution across the formal evaluation funders is relatively proportionate, but again is only of limited relevance, given the overall lack of engagement with this contextual factor.

Figure 5.14 displays the data from the eight contextual variables discussed individually above on a spider diagram, where 0-4 represents the measurement scale, and the average value for evaluations from each governance centre is plotted on each of the eight rays of the diagram.[32] Here, we can see how evaluations funded by actors at the EU level and in Germany appear to resemble each other, while evaluations funded by UK based actors scored lower on the political environment, external events and circumstances, as well as attention to other sectors.

[32] Given that the time horizon was measured on a 6-point scale and all other variables on a 5-point scale (see above), I transformed the time-horizon variable into the same 5-point scale of the other variables.[32]

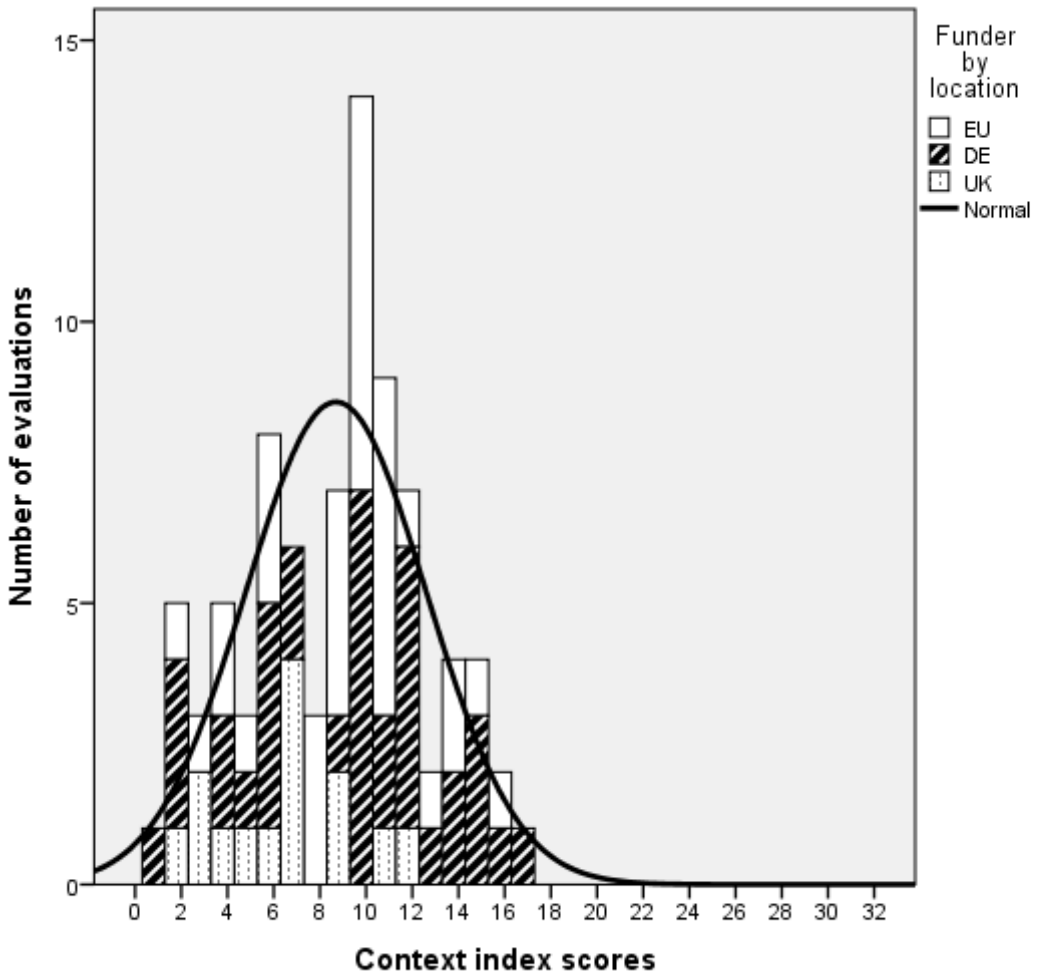
Figure 5.14: Average scores on contextual variables by governance centre



In order to further explore and bring together the eight contextual dimensions, I calculated Pearson correlations among the contextual variables (see Appendix 4). By and large, the variables correlate weakly (i.e., in most cases below 0.3) and insignificantly, indicating that they measure different aspects of contextuality. Therefore, I summed the eight contextual variables in order to generate an overall 'context score' or indicator for each evaluation (I transformed the time variable onto a 0-4 scale, see above). Figure 5.15 presents the data emerging from this process and reveals that no evaluation reached the theoretical minimum (0) or the theoretical maximum (8 x 4 = 32) on the contextual scales. With an overall mean of $M = 8.70$ and a standard deviation of $SD = 3.91$, the distribution clusters on the lower end of the spectrum. The highest score (17) is just over 50% of the theoretical maximum of 32. This means that no formal climate policy evaluation has a particularly large aggregate score on the contextual variables. The added normal distribution curve shows that the overall distribution tends towards a normal distribution, even though some variation remains. These variations also emerge from the different

contributions of evaluations supported by funders in the three governance centres discussed above.

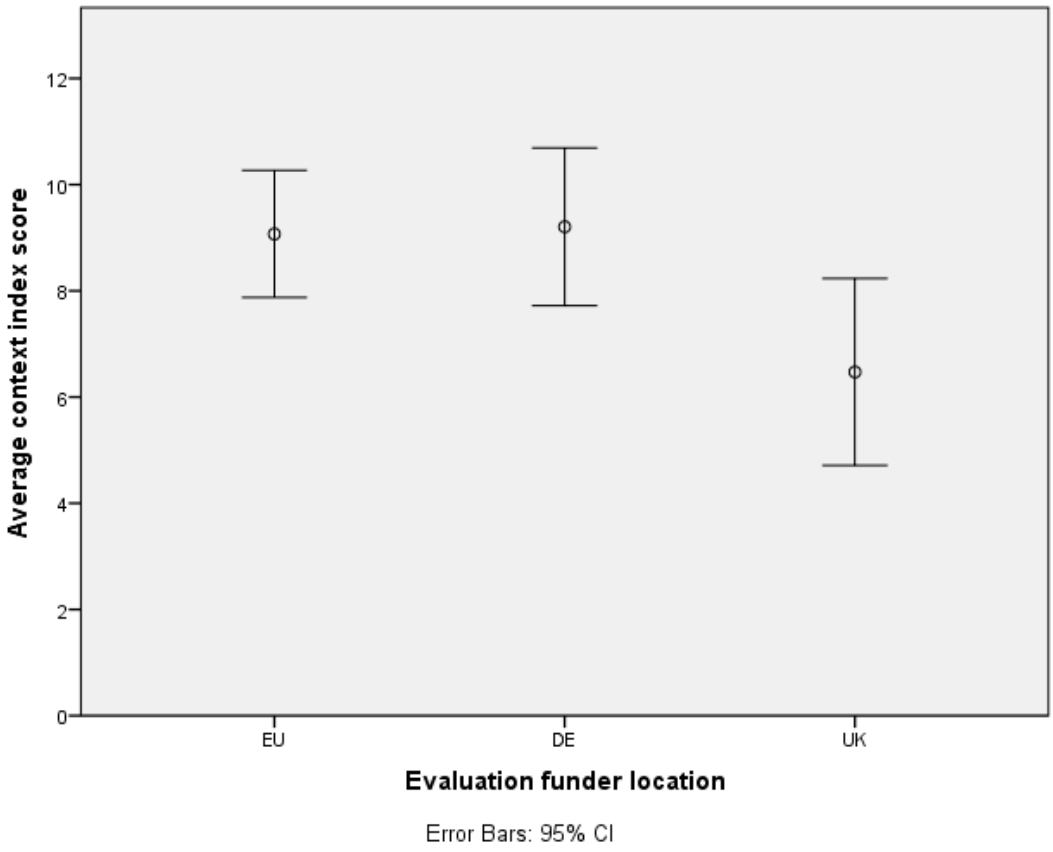
Figure 5.15: Index of contextual variables in formal evaluations



Across all the variables, it becomes noticeable how UK funded evaluations cluster towards the lower end of the spectrum (i.e. they pay less attention to context across all the variables with $M = 6.47$, $SD = 3.05$), compared with the EU-level ($M = 9.07$, $SD = 3.43$) and Germany ($M = 9.21$, $SD = 4.39$), which have both a broader range and contribute significantly to the higher-scoring end of the spectrum (especially Germany). Figure 5.16 presents these mean differences visually, showing the notable difference between the EU level/Germany and the UK, but also some overlap in the confidence intervals. A one-way Analysis of Variance (ANOVA) to compare the three averages by the location of the evaluation funders proved

marginally significant with $F(2, 81) = 2.852$, $p = .064$. Chapters 8 and 9 will pick this back up.

Figure 5.16: Average scores on contextual index by governance centre[33]

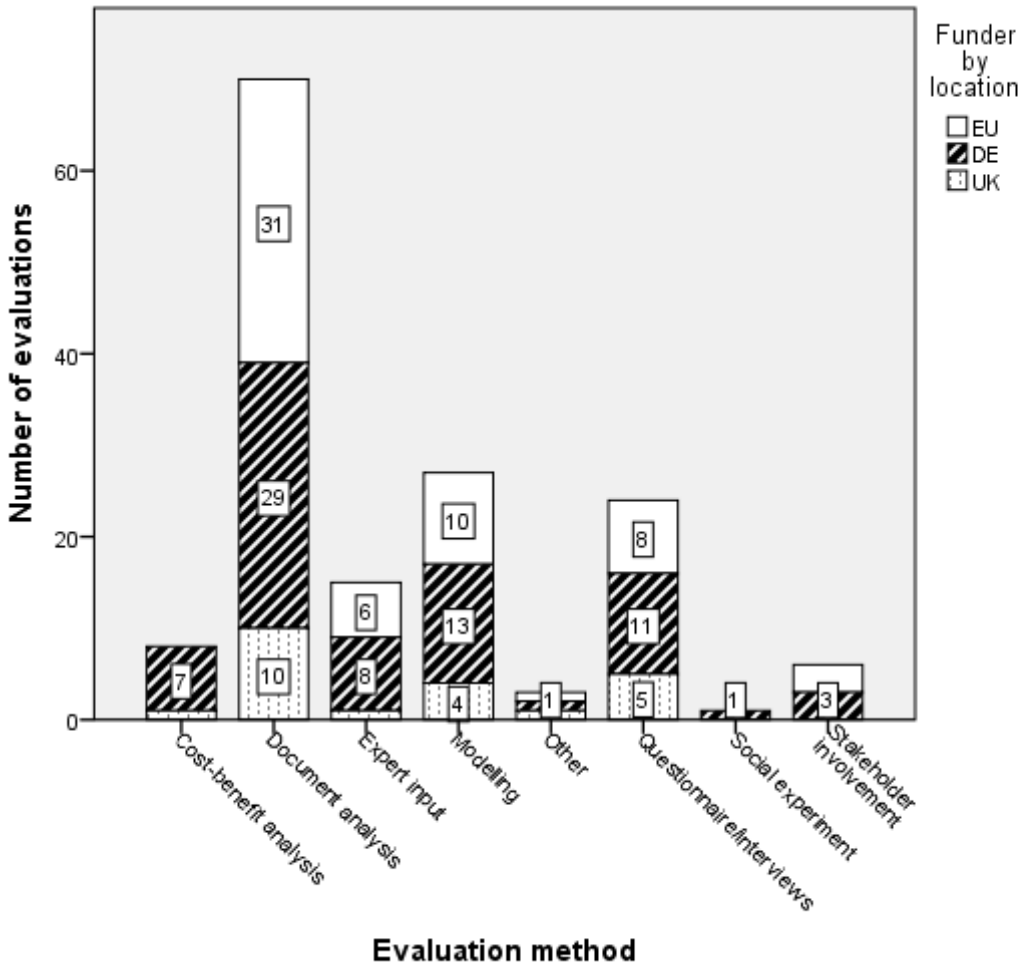


As Chapter 2 has argued, in addition to the variables considered above, there are also other ways in which climate policy evaluations can, in one way or another, take contextual factors into account. One way is through methodological plurality. Using a greater range of methods is one potential way of discerning the possible multifarious effects of a climate policy, given that each method comes with unique strengths, but also potential blind spots (see Rog, 2012). Therefore, Figure 5.17 presents the methodological approaches used in the evaluations. It shows that the

[33] Note that multiple mentions were possible, so that the overall number exceeds the sample of 84 in this figure.

most popular evaluation method is a document analysis (for example a literature review), followed by modelling and questionnaires/interviews—thus in most cases, evaluations drew only on a fairly similar and limited set of methods. Other methods were used to a considerably lesser degree, but across the whole sample, there was a broad spectrum of methods that at least one evaluation used. Methods that incorporate views from stakeholders (such as direct stakeholder involvement or questionnaires and interviews) were only used moderately.

Figure 5.17: Types of methods used in formal evaluations

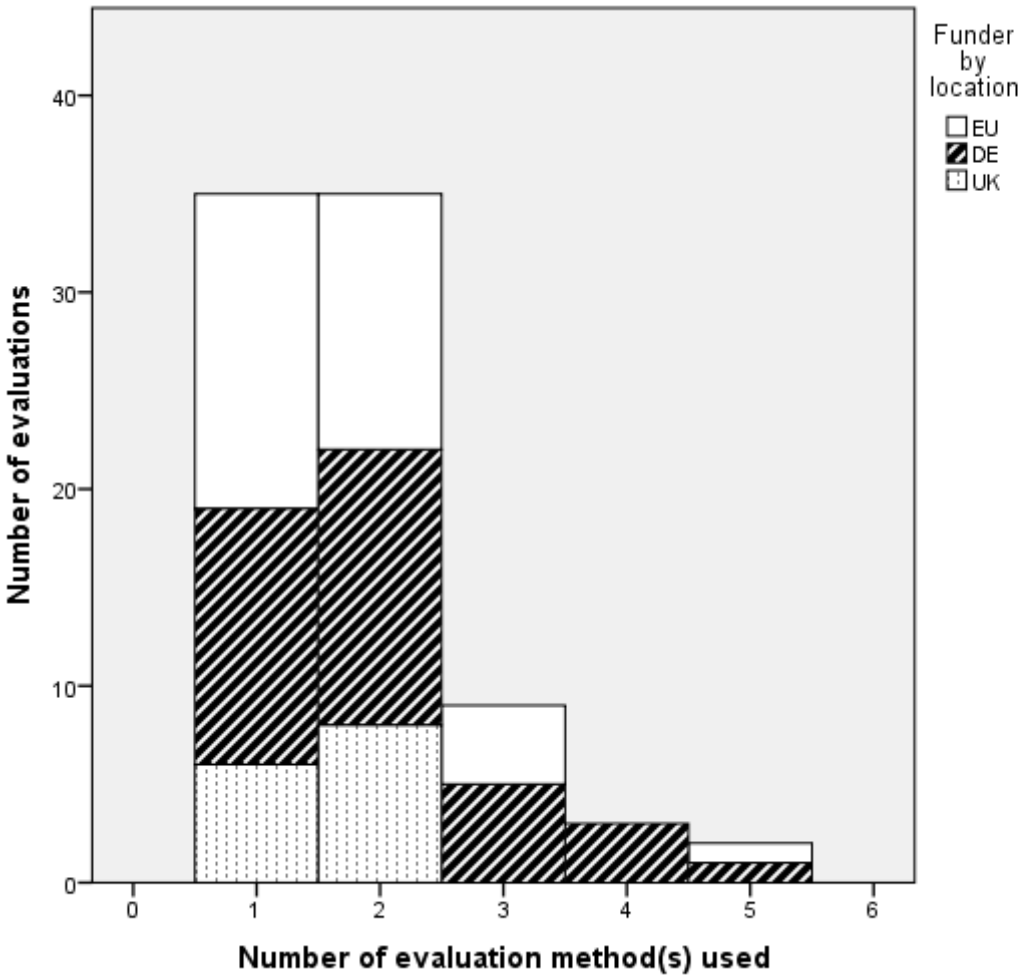


Furthermore, Figure 5.17 shows that there are some differences by the evaluation funder. While funders from Germany supported evaluations with all types of methods (although arguably to different degrees, see the height of the striped bars), EU level actors did not fund any evaluations that used cost-benefit analysis

(CBA) or social experiments. By the same token, UK based evaluation funders did not support any evaluations that used social experiments or stakeholder involvement.

In addition to the type of the evaluation method, it is also relevant how many different methods were used within a single evaluation. As argued above, using more than one method may be an indicator of efforts to capture contextual effects and assess the potentially multiple effects from a single climate policy. Figure 5.18 demonstrates how many evaluations (y-axis) used how many different methods (x-axis). It shows that most formal evaluations use either one or two methods. Evaluations that use three or more methods remain rather rare.

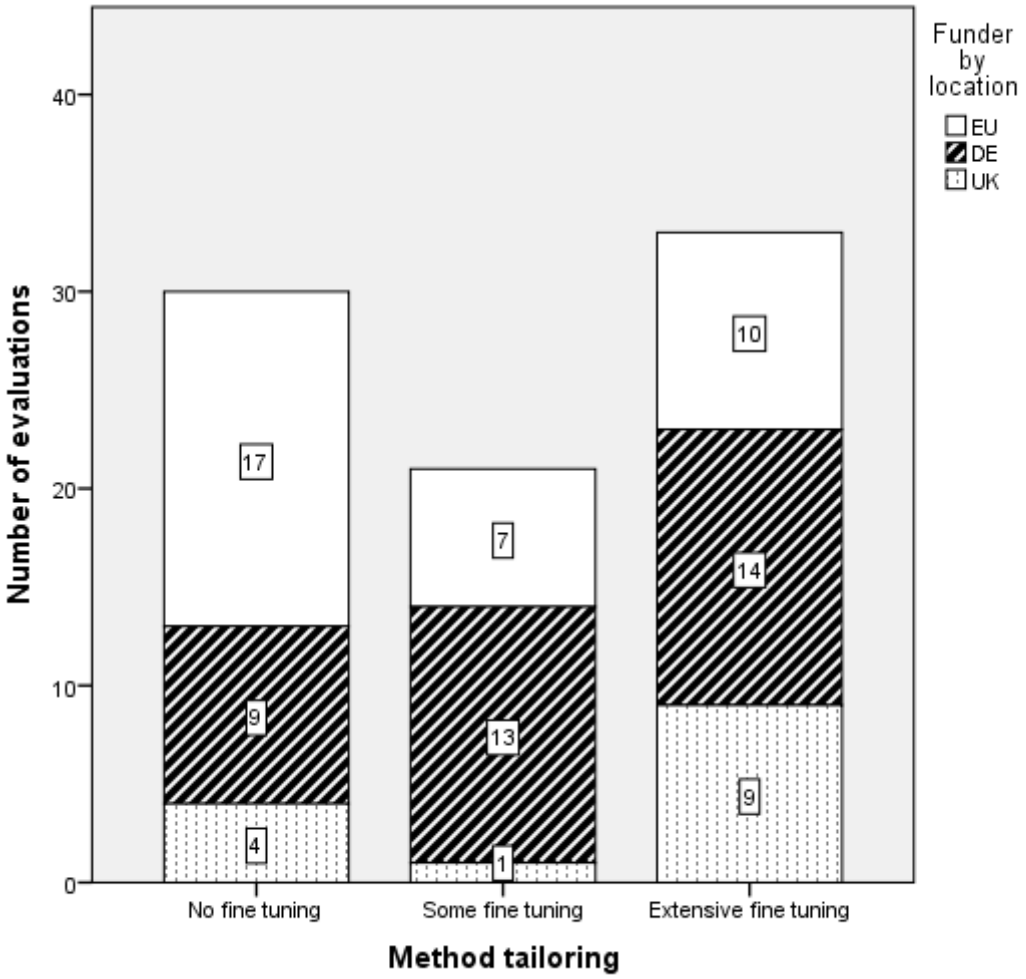
Figure 5.18: Formal evaluations by evaluation method



All evaluations contained in this sample used on average $M = 1.83$ methods ($SD = .93$). There are also differences by funder location. First, considering the different shading in Figure 5.18 reveals that UK based actors did not fund any evaluations that used more than two methods. Actors from Germany and, to a lesser extent, actors based at the EU level mainly funded the evaluations with three to five methods. There are no evaluations that use more than five evaluations. Statistically, this means that evaluations funded by Germany actors use on average the greatest number of methods ($M = 2.03$, $SD = 1.06$), followed evaluations funded by EU level actors ($M = 1.74$, $SD = .90$) and finally the UK ($M = 1.57$, $SD = .51$). However, a one-way ANOVA to compare these three means returned statistically insignificant results [$F(2, 81) = 1.55$, *ns*].

It is not only the type or the number of methods used, but also the extent to which methodological approaches have been tailored to the specific evaluation that matters for the extent to which contextual factors can be taken into account. For example, a survey or a model that has been specifically created or at least calibrated for an evaluation is likely to better fit with the policy and its context than an 'off the shelf' method that is simply applied without much attention to this kind of fit. Therefore, Figure 5.19 depicts to what extent methods used in the evaluations have been calibrated towards the context in question. The data reveal that nearly half of the formally-funded evaluations exhibited extensive tailoring or fine tuning, with a roughly equal number of evaluations whose methods showed no signs of tailoring. Looking at the location of the evaluation funder (shading of the bars) shows a relatively even distribution, although closer inspection shows that UK based evaluation funders financed a disproportionately large number of evaluations with extensive methodological fine-tuning, which are thus more context specific.

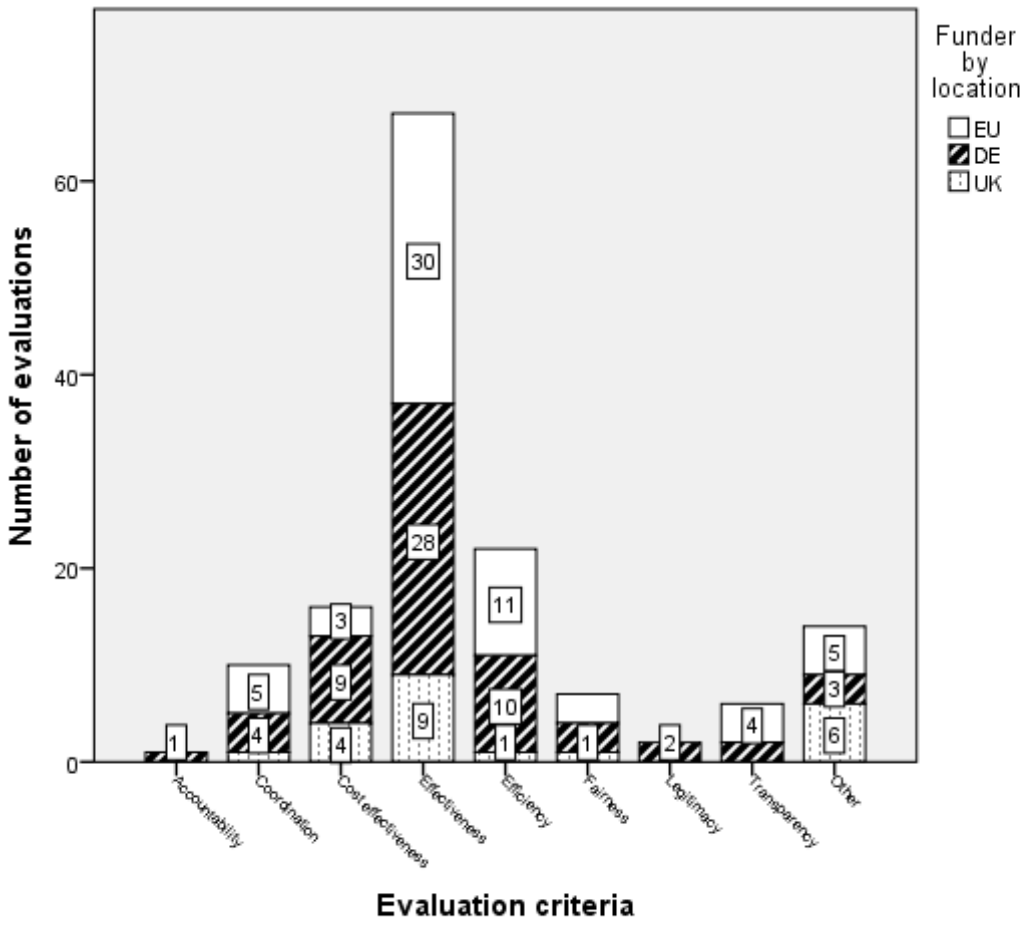
Figure 5.19: Methodological ‘tailoring’ in formal evaluations



Another important contextual element concerns the evaluation criteria that a climate policy evaluation uses. Analogous to the methodological argument, a plurality of criteria could, again, point to greater attention to context (because evaluation with different criteria may pick up different kinds of policy effects). Figure 5.20 shows the type of evaluation criteria, and the number of evaluations that used them (multiple mentions were possible). Policy effectiveness (with a view to its goal attainment) is by far the most widely-used evaluation criterion, followed by efficiency, and cost effectiveness. Notably, accountability and legitimacy were hardly used at all. Looking at the evaluation funders reveals no clear trends across the data, although UK based funders appear to support evaluations with a particular focus on effectiveness and cost effectiveness, whereas particularly Germany based actors funded nearly all of the side effect evaluations. Altogether formally-funded

evaluations have used a very broad spectrum of criteria, but the majority concentrates on just a few of them.

Figure 5.20: Types of criteria used in formal evaluations[34]

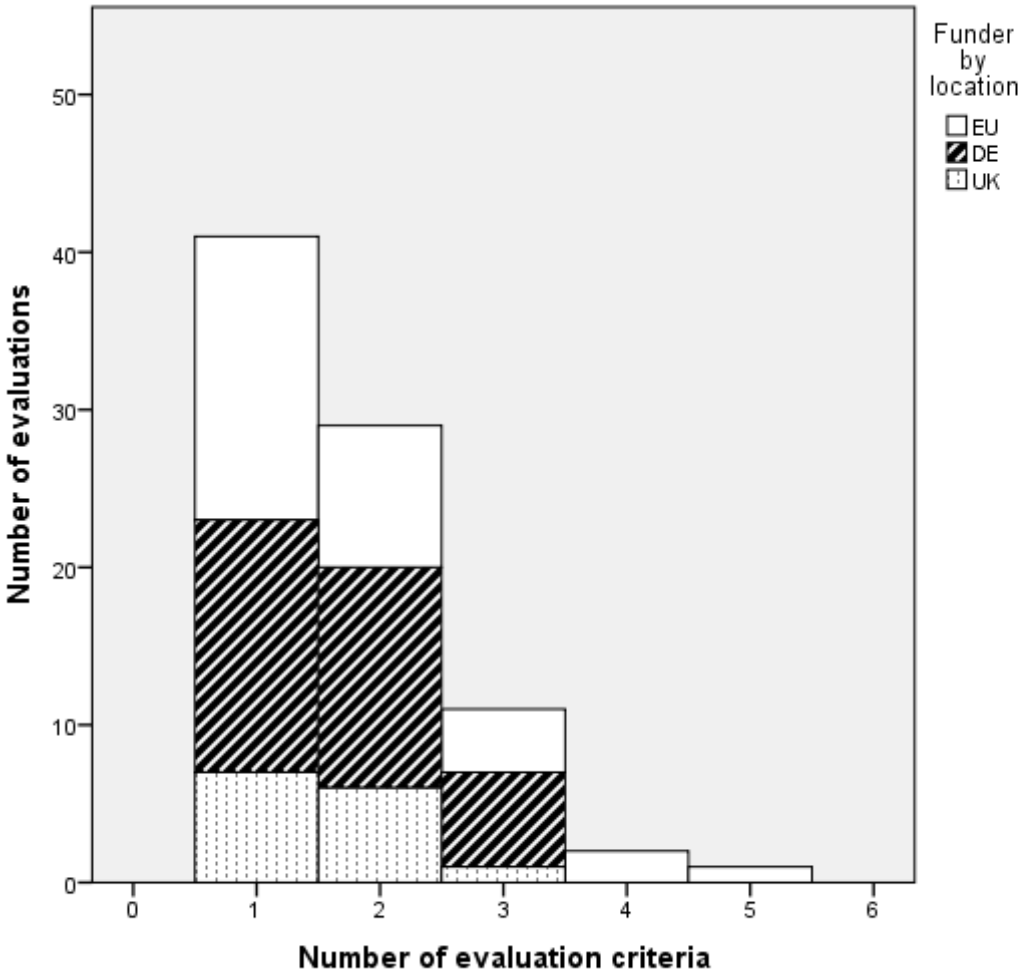


Similar to the discussion on evaluation methods above, it is also relevant to consider the number of criteria used in the evaluations. Figure 5.21 shows that most formal evaluations used mainly one or two criteria, and very few used more. On average, formal evaluations contained *M* = 1.73 (*SD* = .87) criteria. Looking at the evaluation funder locations, evaluations funded by EU level actors used *M* = 1.79 (*SD* = 1.07), with slightly more evaluation criteria in evaluations supported by Germany based funders (*M* = 1.72, *SD* = .74) than in evaluations supported by UK-

[34] Multiple mentions possible.

based funders (M = 1.57, SD = .65). However, a one-way ANOVA revealed that these differences were not statistically significant with [$F(2, 81) = .32$, *ns*]. It should be noted, however, that UK based actors funded only a single evaluation that contains more than two criteria. Thus, even though the spectrum of criteria used may be wide, in practice most formally-funded evaluations only focused on one or two criteria.

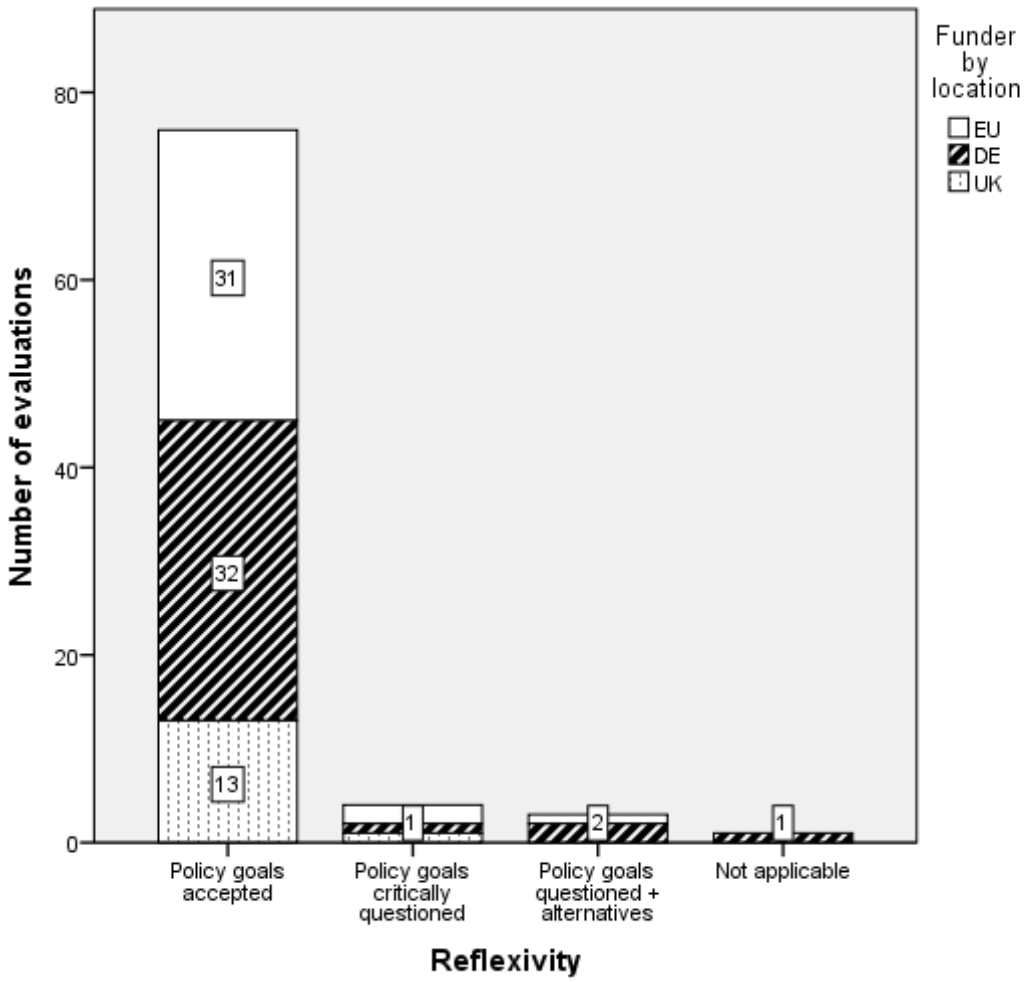
Figure 5.21: Number of criteria used in formal evaluations



Connected with criteria are the policy's guiding goals, which may frame the evaluations. Figure 5.22 considers to what extent climate policy evaluations are 'reflexive,' (see Huitema et al., 2011) or in other words, how many evaluations either take extant climate policy targets as a given (no reflexivity) and evaluate against them, or question policy targets critically and even offer alternatives. It reveals that,

by and large, formally-funded evaluations are not very reflexive. Almost all evaluations accepted given policy targets as a given and did not engage with them critically. This is important because as Chapter 2 argues, shifts in context could impact on the relevance of a target set when the policy was first implemented or decided upon. Only 9.52% of the evaluations – mainly funded by actors based in Germany and at the EU level - engaged critically with prevailing policy goals.

Figure 5.22: Reflexivity in formal evaluations



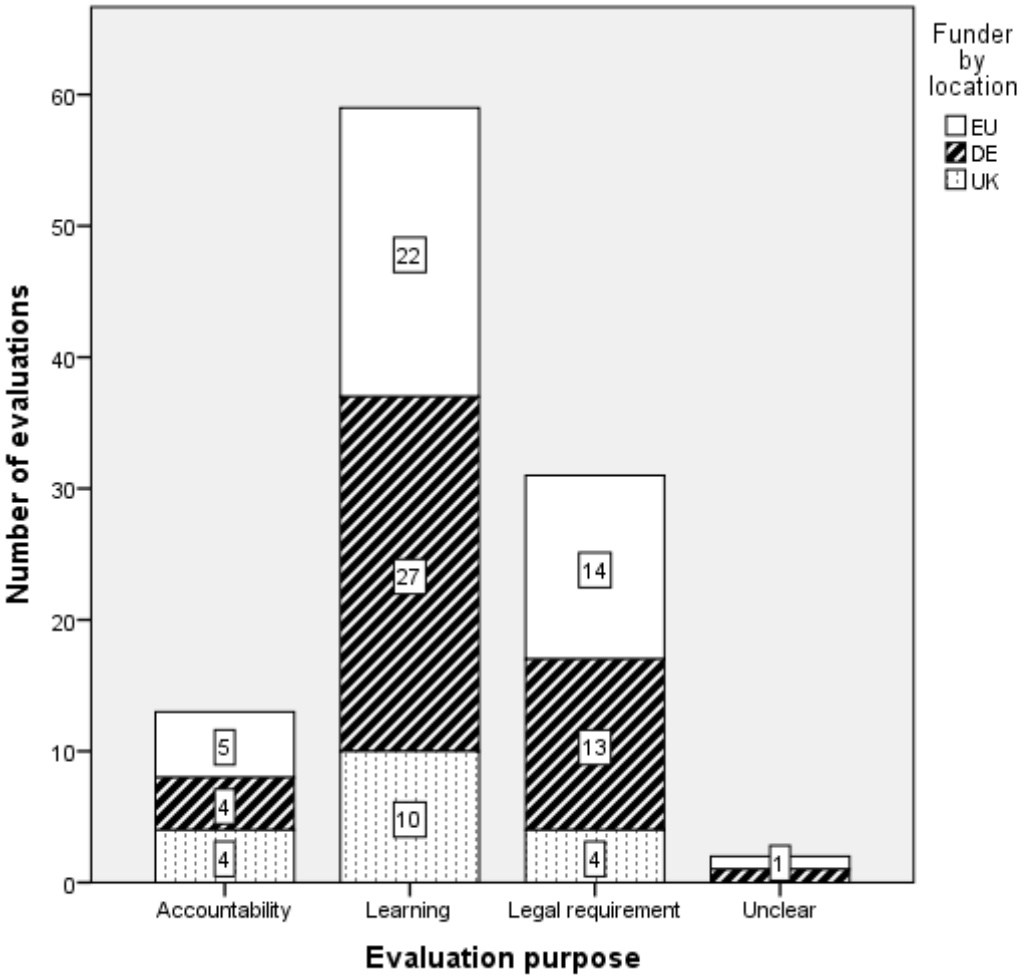
- There are, in sum, a number of different ways to empirically assess attention to contextual factors in evaluations.
- The indicators assessed here do not all point in the same direction for formally-funded climate policy evaluations.

- However, across the different – and in many ways unique – dimensions, it is clear that there is generally a limited treatment of contextual factors by formally-funded evaluations.
- This analysis also indicates differences across the governance centres of the evaluation funders: evaluations funded by actors based in Germany tend to account for context more than evaluations funded by EU-level actors, with evaluations funded by UK-based actors as the least context-sensitive.
- Formal evaluations use, in general, few methods and criteria and are not very reflexive.

5.4 Interaction

This section considers the extent to which the climate policy evaluations reveal interactions between governance centres, which is one of the key postulates of polycentric governance (see Chapter 2). The first aspect in this section in Figure 5.23 considers the stated or overt evaluation purpose (for a more detailed discussion, see Chapter 4). From a polycentric perspective, it matters whether an evaluation simply responds to a legal requirement or whether it was explicitly conducted in order to foster learning, perhaps even with a view to providing lessons for other governance centres. Figure 5.23 reveals that by far the most widely identified evaluation purpose was learning, followed by legal requirements and to a lesser extent accountability. Furthermore, there is a relatively even distribution of evaluations funded by actors from the EU level, from Germany and from the UK in each category.

Figure 5.23: Evaluation purpose[35]

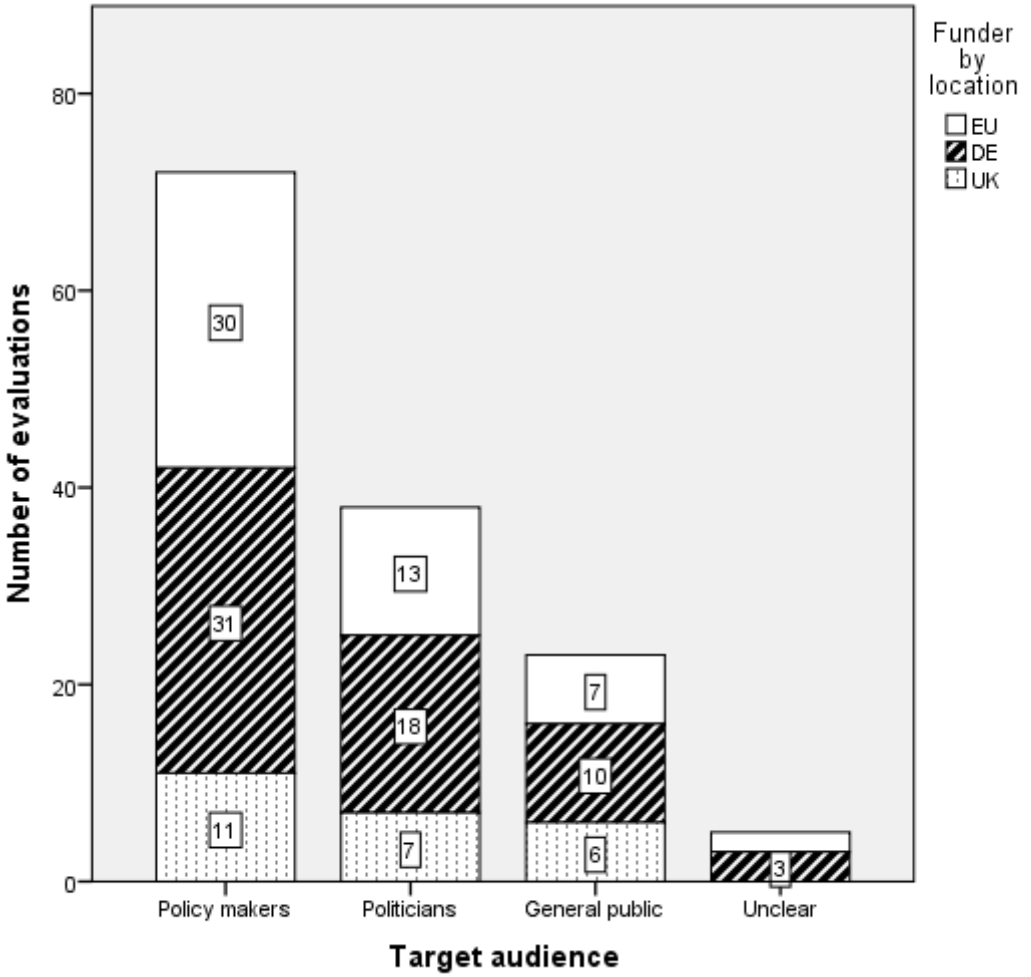


In addition to the (stated) impetus for the evaluation, with a view to interacting governance centres it also matter for which audience the evaluation was (explicitly) generated. Figure 5.24 presents the number of formal evaluations by target audience (multiple mentions were possible). It reveals that almost all (87.80%) formal evaluations were geared towards policy-makers, which would for example be ministry employees or other staff working on more detailed aspects of policy. Given that policy-makers were also among the biggest category of funders, they are evidently mainly funding evaluations for themselves. This is followed by politicians

[35] Multiple responses possible.

(45.24%) and the general public (27.38). Evaluation funders from the EU level, from Germany and from the UK contributed by and large proportionately to each category.

Figure 5.24: Target audience[36]

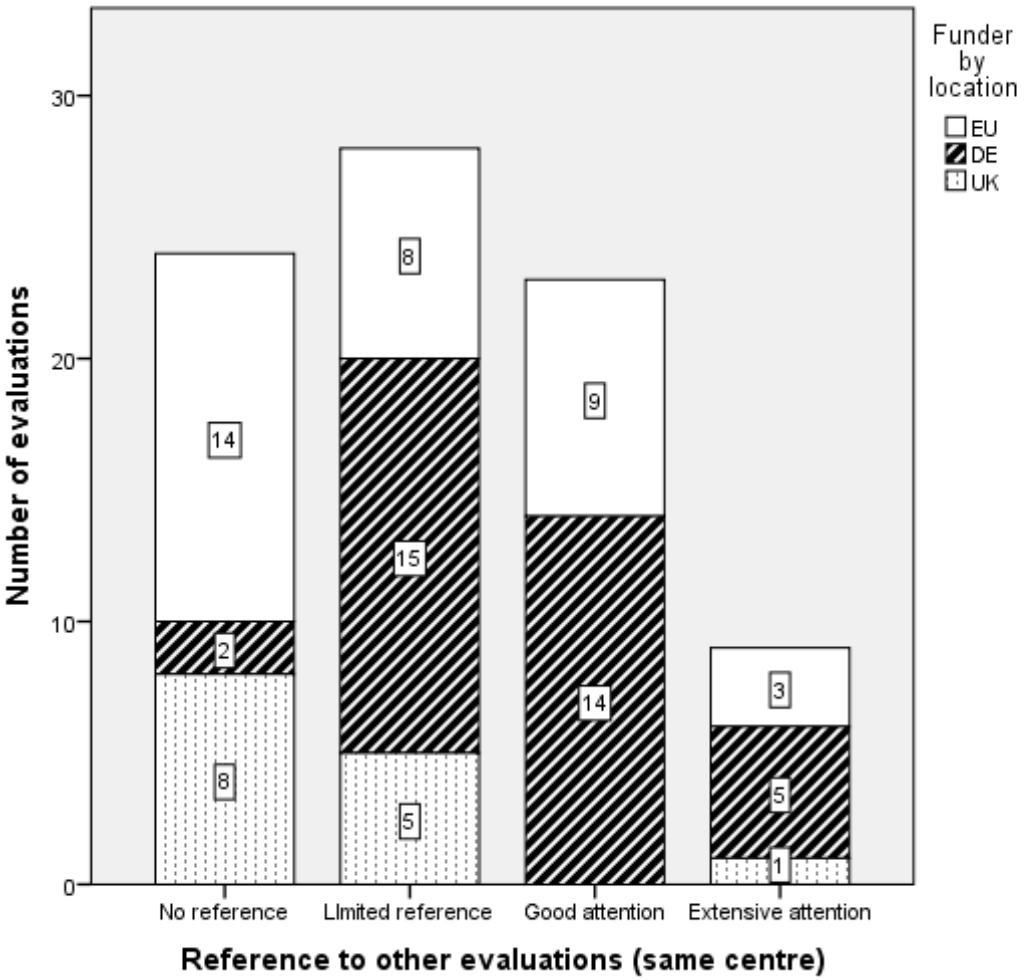


Another way of considering interacting governance centres is to understand to what extent an evaluation draws on other evaluations – evaluations that focus on the same centre, but not necessarily the same policy – in order to understand the effects of climate policy. This is perhaps the most immediate and direct type of interaction that can be detected from analysing evaluations, where interaction between governance centres manifests directly through the evaluation. Figure 5.25 reveals that information for formal climate policy evaluations. The bar graph shows that

[36] Multiple mentions possible.

61.90% of the evaluations made either no reference, or limited references, to other evaluations focusing on the same centre. However, 38.10% of the formal evaluations paid good or even extensive attention to other evaluations focusing on the same centre (manifest through citing findings or insights from other evaluations), thus tapping into a wider web of knowledge on the policy in question than that generated directly by the evaluation. The differently-shaded bars show that there was some variation depending on the location of the evaluation funders, indicating that UK based actors tended to fund more evaluations with no or limited attention to studies about the same governance centres than funders from Germany or the EU level.

Figure 5.25: References to other evaluations focusing on the same centre



Analogously, Figure 5.26 expresses the extent to which formal evaluations incorporated insights from or focused on other governance centres. An example may

be an evaluation of the EU Emissions Trading System, which makes references to the experience in the USA. Figure 5.26 shows that attention to other governance centres within formal evaluations is even more limited. Only 16.67% of the evaluations paid good or extensive attention to experiences in other governance centres – overall, this points to limited interaction between governance centres vis-à-vis formal climate policy evaluations. Distinctions by evaluation funders were such that EU level funders showed a slightly higher propensity to fund evaluations that paid more attention to other governance centres compared to those funded by actors in Germany and, especially, funders from the UK. Looking across Figure 5.25 and Figure 5.26, and compared to the common practice of citing others' work in academia, it is noticeable how self-referential and insular evaluations in this sample tend to be.

Figure 5.26: References to evaluations focusing on other centres

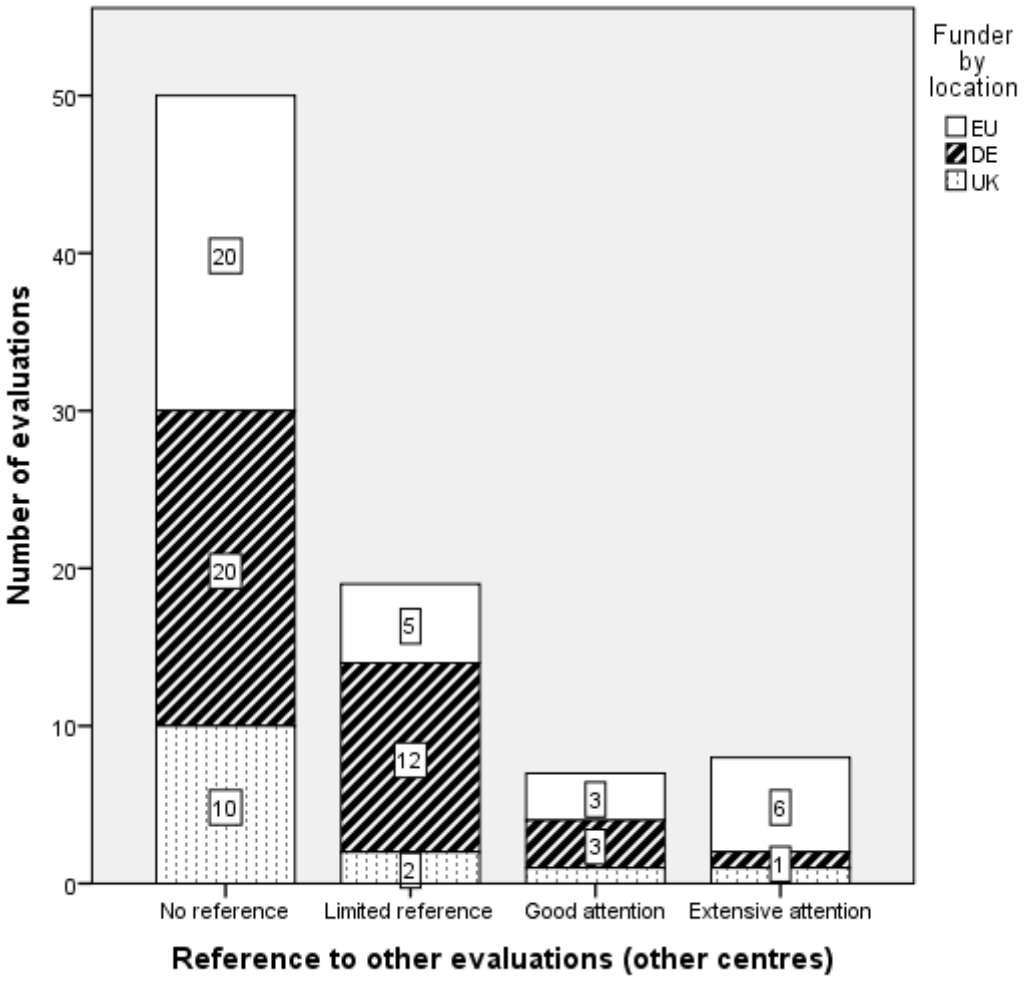
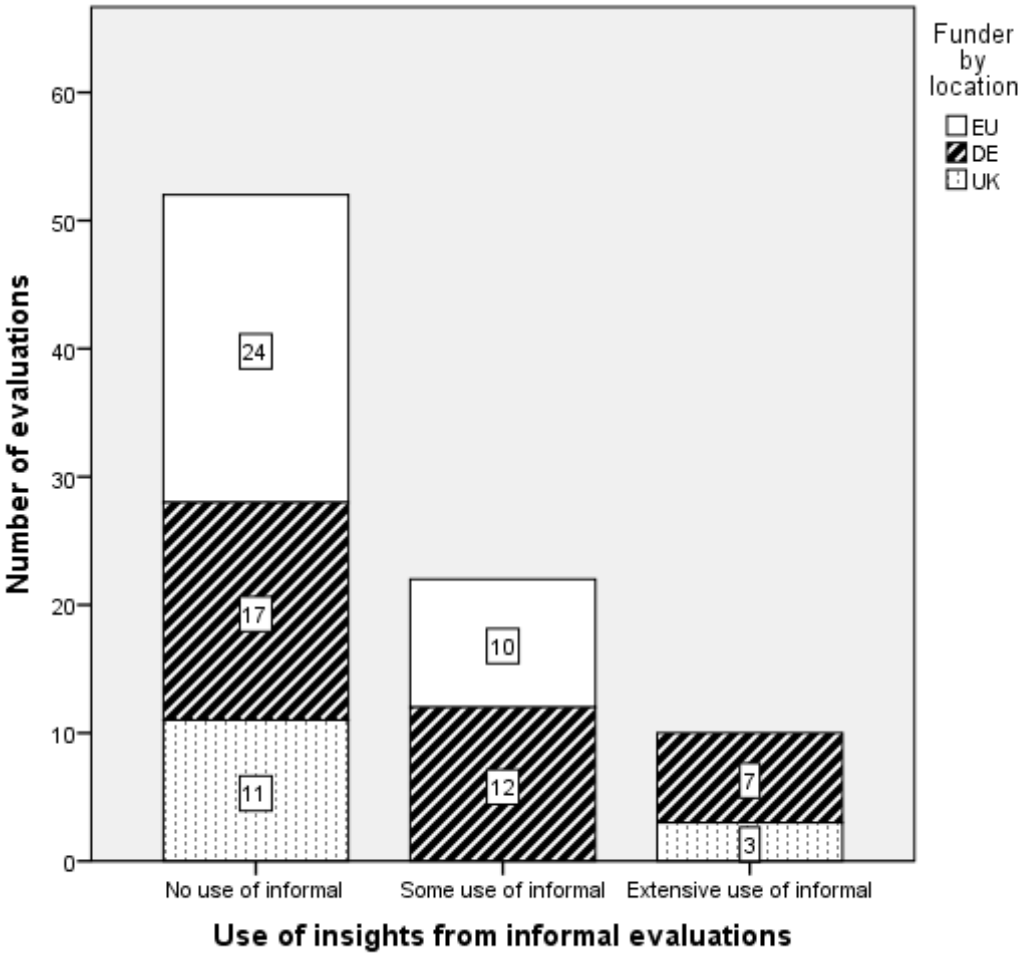


Figure 5.27 then turns to the extent to which the formal evaluations in our sample draw on insights or data from informal evaluations, or to what extent interactions between these different types of evaluation there are. It demonstrates that a strong majority of 61.90% of formal evaluations (first bar) did not make any reference whatsoever to informal evaluations. 26.19% of the formal evaluations made some reference to informal evaluations (the second bar), but only 11.90% of formal evaluations extensively referenced or used data from informal evaluations. Looking at the distribution by evaluation funders (the shading of the bars) shows that only actors based in Germany or the UK funded evaluations that made extensive reference to informal evaluations.

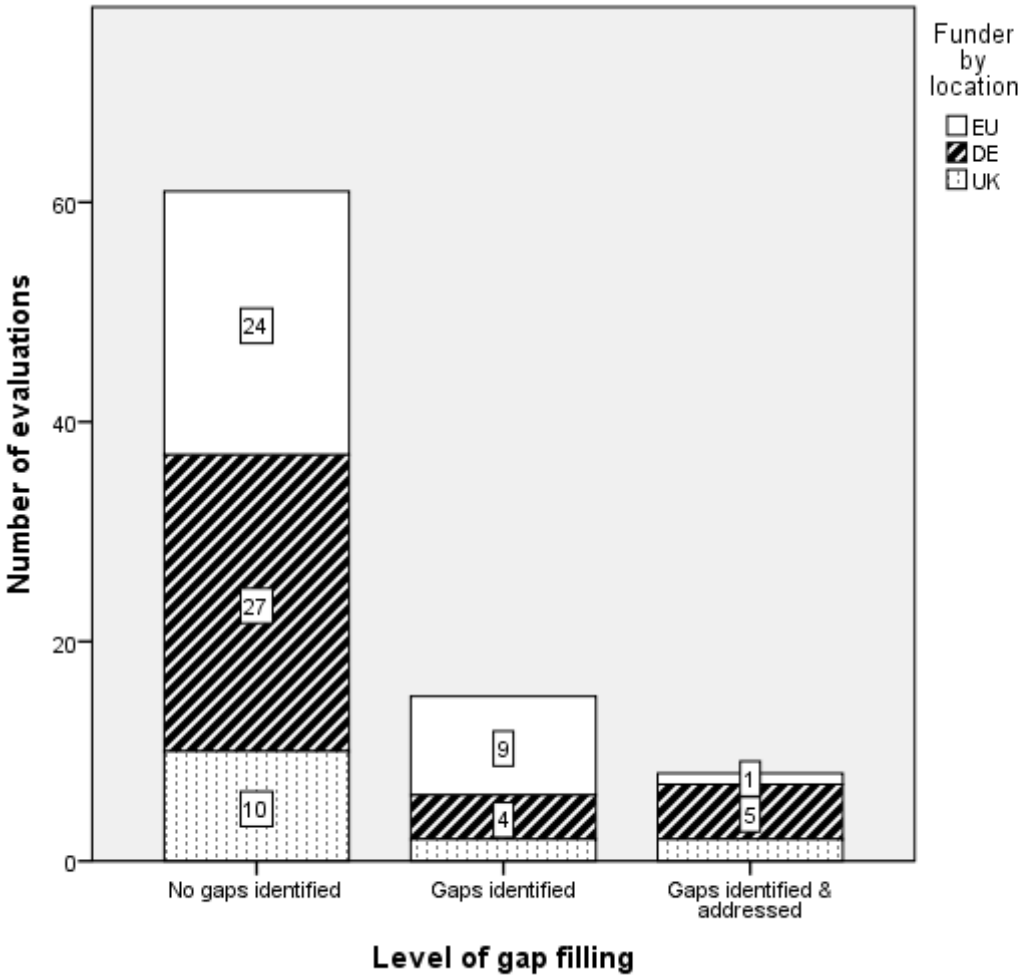
Figure 5.27: Use of insights from informal evaluations



Furthermore, Figure 5.28 demonstrates data on the extent to which formal evaluations attempted to spot and address any gaps left by informal evaluations. It

shows that 72.62% of all formal evaluations made no explicit attempts to address gaps left by informal evaluations (first bar). Fully 17.86% of the formal evaluations spotted gaps in informal evaluations, but only 9.52% of the evaluations spotted gaps *and* addressed them. Considering the data by the location of the evaluation funders (the shading of the bars) demonstrates that while formal actors at the EU level were particularly strong in funding evaluations that spotted gaps, actors in Germany and in the UK were by and large the only ones that funded evaluations that spotted *and* addressed gaps.

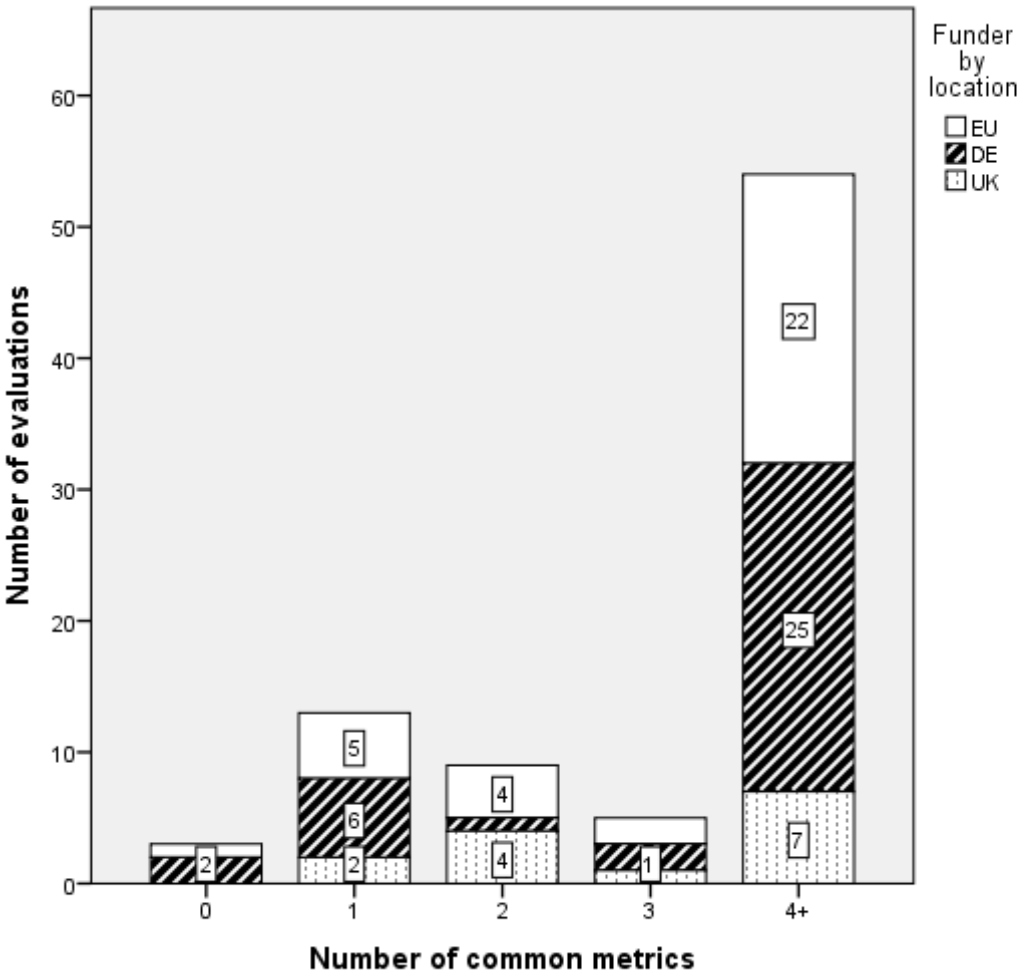
Figure 5.28: Formal evaluations filling gaps in informal evaluation



Another way of enabling interaction between governance centres via evaluation is to produce quantitative comparability metrics. These may, for example, be the greenhouse gas emissions that a climate policy reduces or the costs that a

policy generates (see Chapter 4). Figure 5.29 depicts the number of quantitative comparability metrics contained in the formal climate policy evaluations whose analysis this chapter presents. It reveals that 64.29% of formally-funded climate policy evaluations contain four or more comparability metrics. In other words, there is a significant quantitative focus in formal evaluations. The distribution by evaluation funder is relatively proportionate (keeping in mind the overall distribution noted in the introductory section).

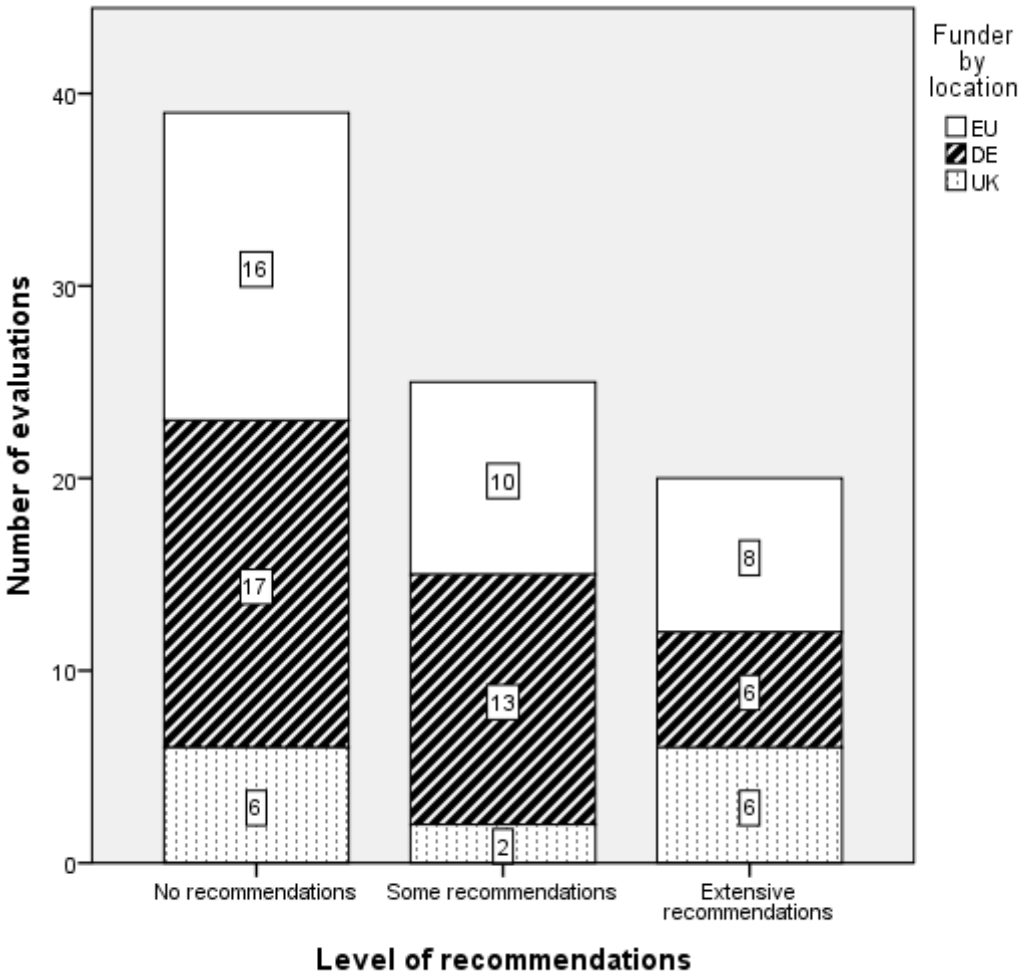
Figure 5.29: Number of comparability metrics in formal evaluations



Another way of stimulating interactions between governance centres and potentially learning or other forms of interactions is through the production of policy recommendations. Figure 5.30 summarizes the extent to which formal evaluations contain policy recommendations. Using the operational definition of an evaluation

discussed in Chapter 4, all evaluations had to include some recommendations or at least a policy-relevant conclusion. Figure 5.30 focuses on specifically stated recommendations in the evaluation. It shows that 46.43% of the formal climate policy evaluations contained no explicit policy recommendations, and only 23.81% of the evaluations contained extensive recommendations. Looking at these data by evaluation funder (the shading of the bars in Figure 5.30) shows that funders at the EU level, in Germany, and in the UK contributed about proportionately to each category, although UK based actors tended to fund a somewhat higher number of evaluations with extensive recommendations.

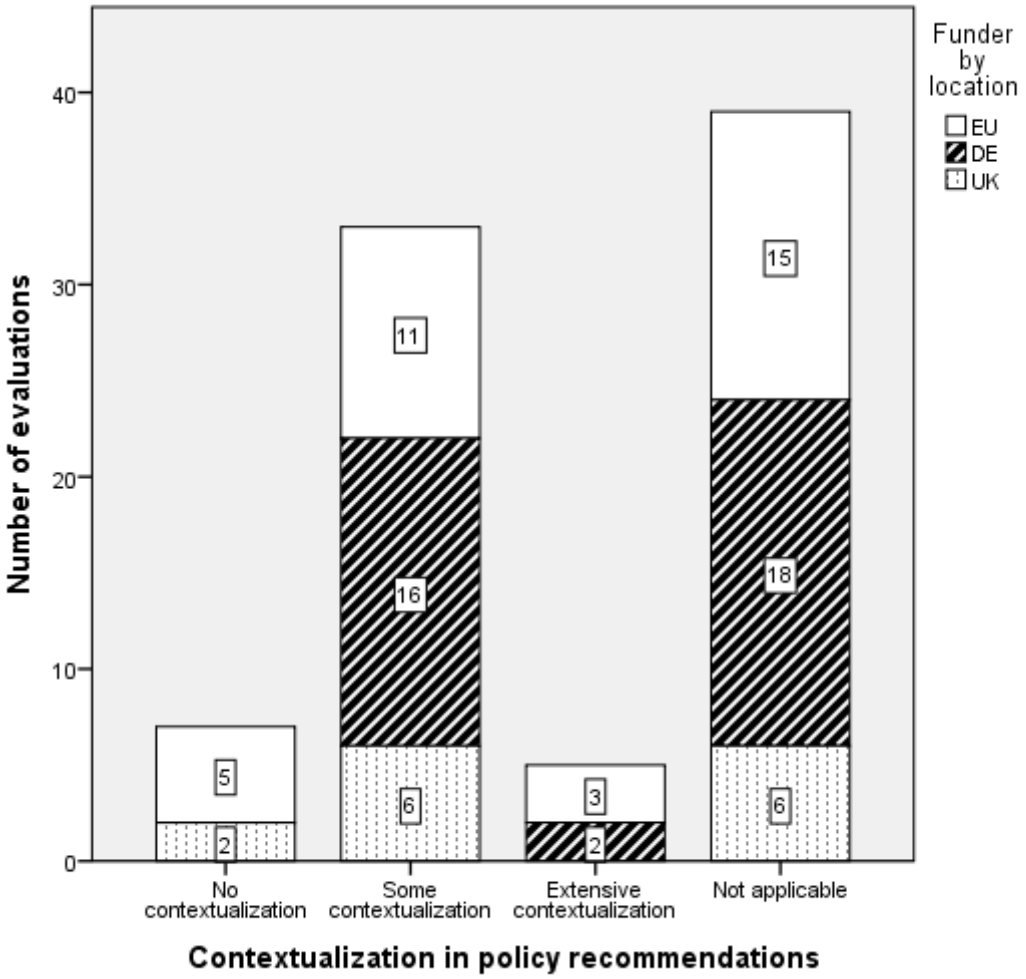
Figure 5.30: Recommendations in formal evaluations



In line with the argument that the context affects how a policy fares (see above and Chapter 2), it also matters that recommendations contain some level of

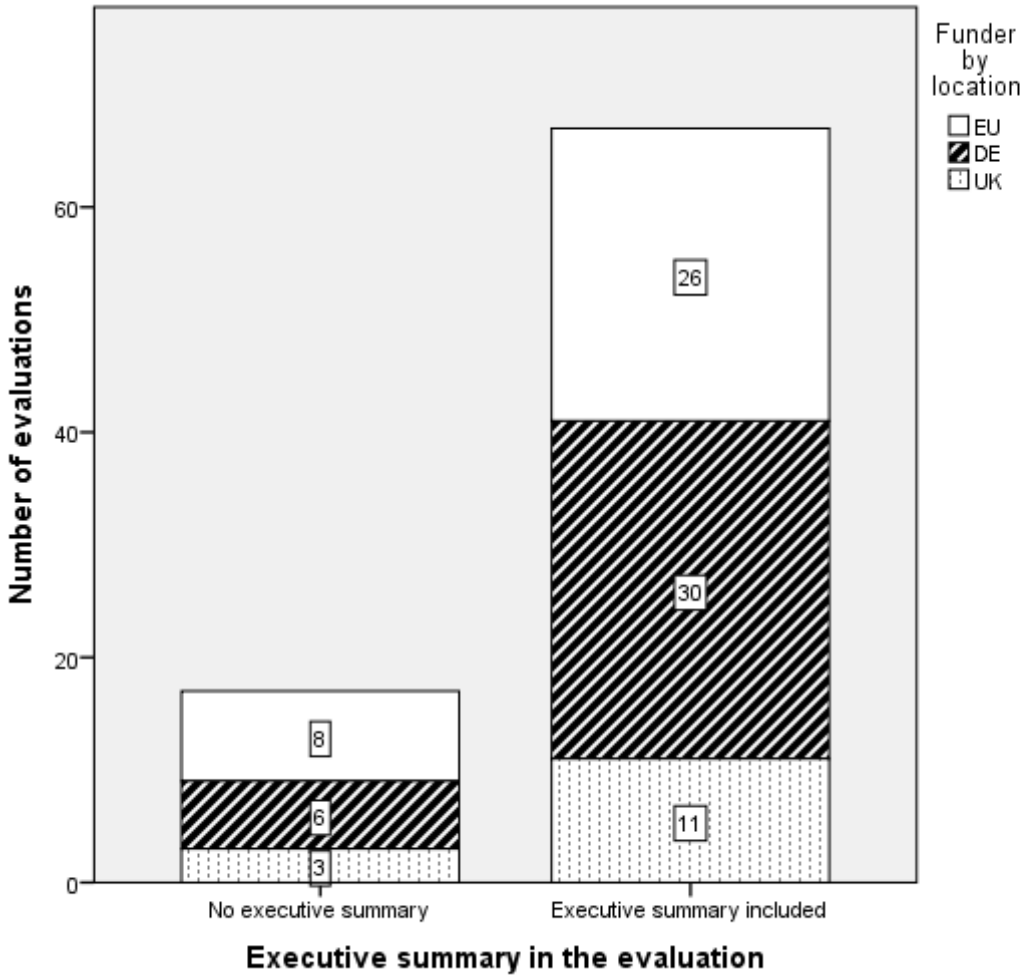
contextual information in order to gauge the extent to which lessons may (or may not) carry from one context to another. For example, if recommendations include information on the political system for which they may be suitable (or from which political system they are derived), this information may impact on the usability of these recommendations across different governance contexts. Figure 5.31 therefore summarizes data on the extent to which the recommendations in formally-funded climate policy evaluations contain contextual information. The considerable 'not applicable' category in the figure corresponds with evaluations that contained no recommendations (see Figure 5.30 above). For the evaluations that contained recommendations, it is noticeable that most of them contained some level of contextualization, but more in-depth, extensive contextualization remains an exception rather than a norm. Looking at the evaluation funders in Figure 5.31 shows that funders from Germany only funded evaluations with some contextualization. By contrast, the UK funded no evaluations that extensively contextualized their recommendations.

Figure 5.31: Contextualization of policy recommendations



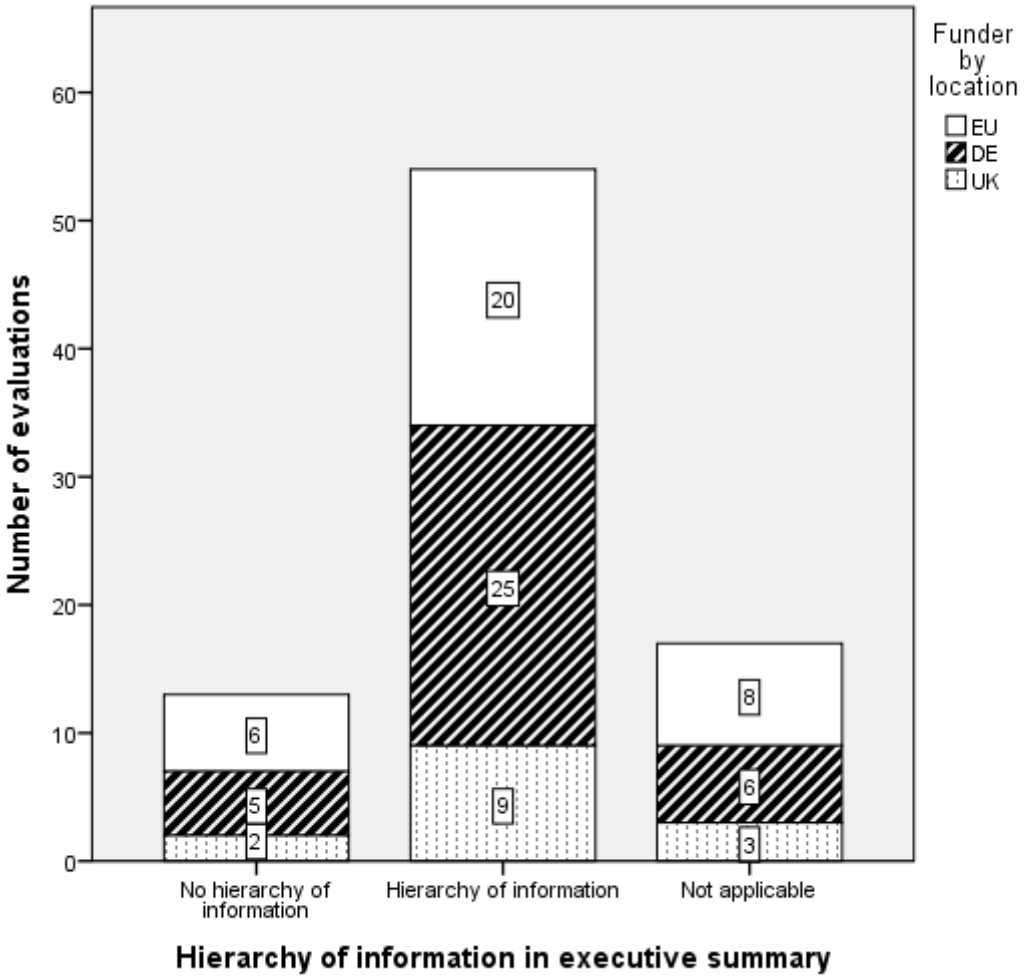
In busy and often time-pressured policy-making situations, shorter and well-structured evaluations may be more impactful than lengthy ones with little indication of the most relevant findings and points (see Chapter 2 and especially Zwaan et al., 2016). Therefore, executive summaries may be a key element in this process because they typically aim to summarize the most salient findings from an evaluation. Figure 5.32 reveals the number of formal evaluations that contained executive summaries. It reveals that 79.76% of the evaluations contains executive summaries, but just under a quarter of the evaluations does not. Funders based at the EU level, in Germany, and in the UK generally supported a proportionate number of evaluations in each category.

Figure 5.32: Executive summaries in formal evaluations



The next Figure 5.33 further unpacks the executive summaries by considering to what extent their structure highlights salient points. As explained above and in Chapter 2, a clear structure in executive summaries such as bullet points, a table or bolding may aid busy policy-makers to quickly discern relevant information. Figure 5.33 shows that 64.29% of the executive summaries contain an internal structure in order to provide some hierarchy of information. In other words, the authors of most formal evaluations took some care to make their evaluations accessible to busy policy-makers and others with an interest in the findings—individuals or institutions who may potentially also be located in other governance centres. Evaluation funders at the EU level, in Germany and in the UK supported evaluations in each category to a proportionate extent.

Figure 5.33: Hierarchy of information in executive summaries



Particularly in multi-lingual European environments, and if evaluative information is to travel well within and potentially beyond Europe's borders, evaluative information in different languages influences to what extent recommendations and lessons can travel across linguistic borders. Figure 5.34 thus depicts the extent to which formal evaluations contained summaries or even whole versions in other languages (than the original language of the evaluation). As Figure 5.34 reveals, providing summaries in other languages remains a relatively rare exception in the evaluations coded for this chapter (14.29%), again pointing to the fact that many formally-funded evaluations appear rather self-referential. Noticeably, evaluation funders in the UK did not fund any evaluations that provided any translations into other languages, an activity that was solely supported by funders based at the EU level and in Germany.

Figure 5.34: Linguistic access to evaluation by funder location

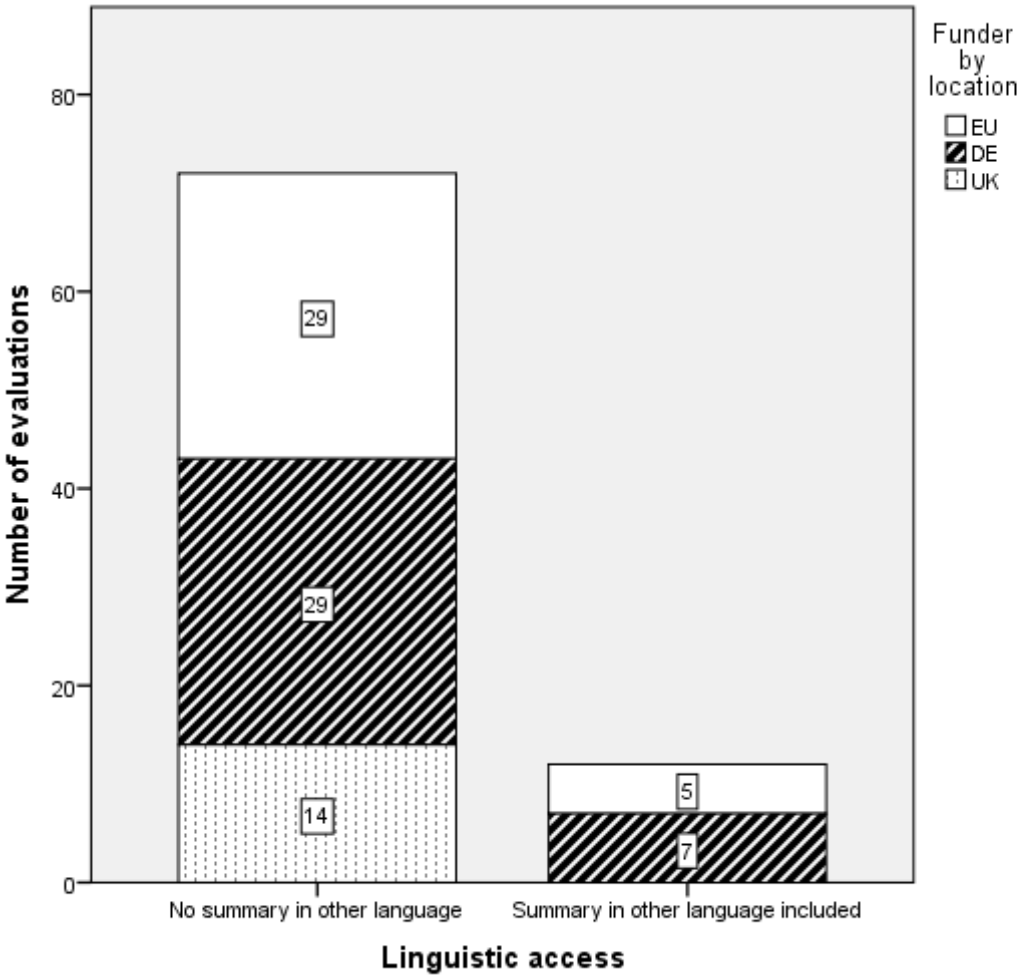
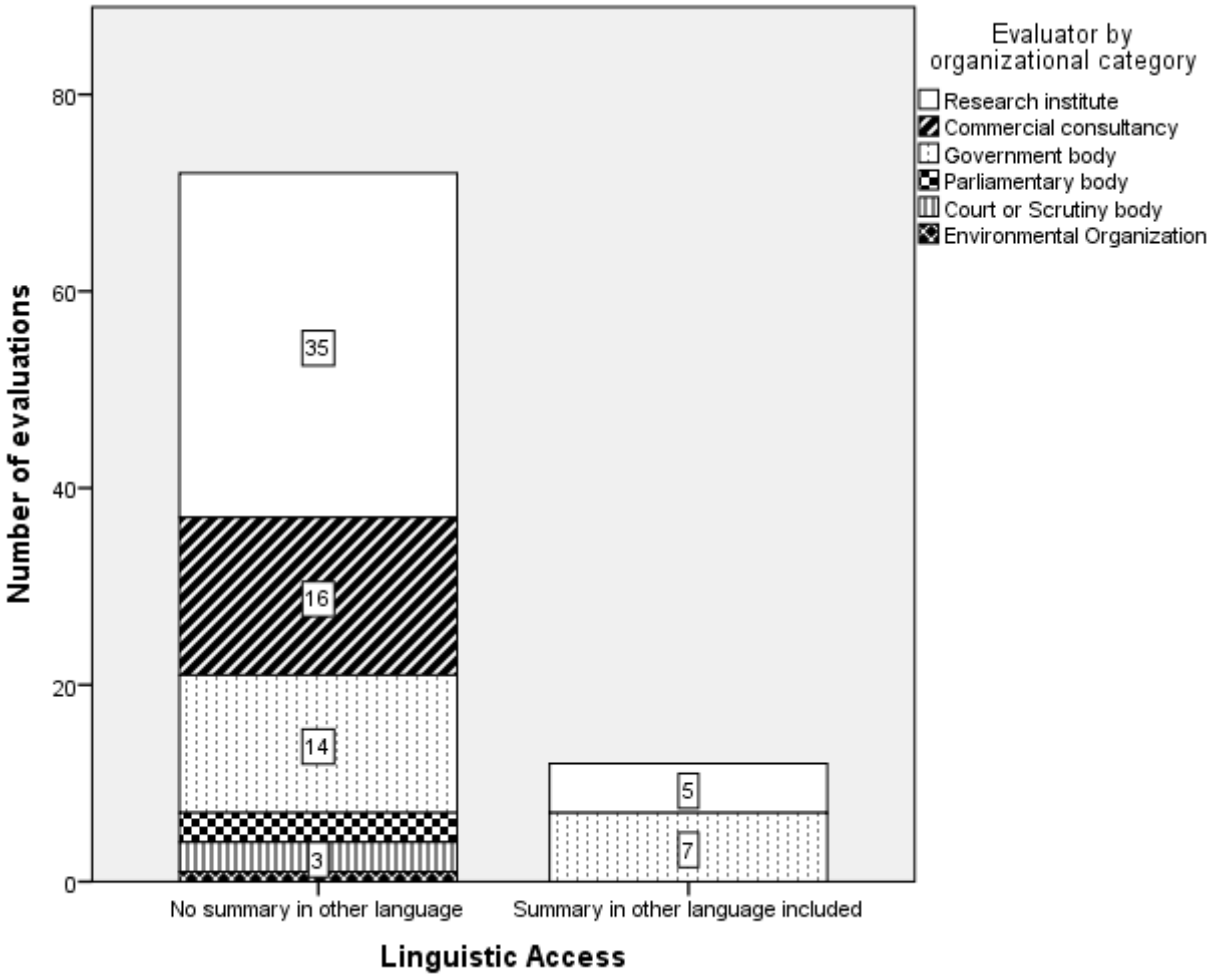


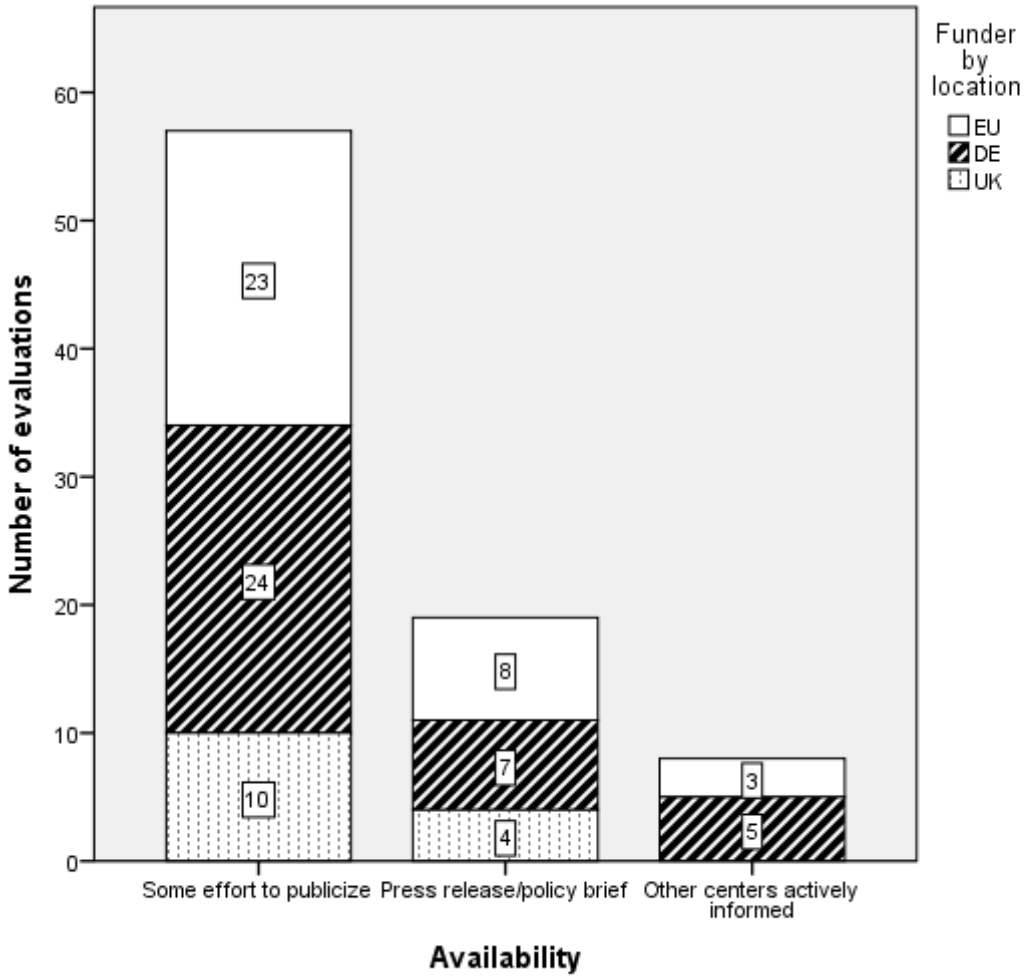
Figure 5.35 presents the same data on linguistic availability by the organizational category of the evaluator. Here, it is readily visible that only research institutes and government bodies provided summaries in other languages. All other evaluators, such as commercial consultancies, parliamentary bodies, courts and scrutiny bodies or environmental organizations did not include translations with their evaluations.

Figure 5.35: Linguistic access by evaluator organizational category



Finally, Figure 5.36 depicts the availability of the formal evaluations, including the effort made to publicize the evaluation and especially the findings (such as for example making the evaluation available on a website, or holding a press conference), which is another important way in which governance centres may interact vis-à-vis evaluation. The bar chart reveals that the authors of 79.76% of the formal evaluations made some effort to publicize their evaluations, but very active communication between governance centres (such as workshops or conferences to share findings) remains very rare (9.52%). Importantly, evaluation funders in the UK did not fund any evaluations that actively informed other governance centres. Combined with the linguistic aspects in Figure 5.34, this suggests that evaluations funded by actors in the UK appear more inward-looking than those funded by actors at the EU level or in Germany.

Figure 5.36: Evaluation availability



- As with the foundational ideas of self-organization and context, this section on interaction demonstrates that there are multiple ways to understand and empirically operationalize interactions between governance centres.
- Formal climate policy evaluations are at best to a limited extent placed to foster interactions between governance centres. Neither their characteristics, nor their evaluation approaches or the publicity done in order to draw attention to the evaluation findings point to an outstanding propensity to foster interactions.
- But it is also important to recognize that a smaller number of evaluations score comparatively highly on several relevant dimensions. Overall and on multiple counts, evaluations funded by actors in the UK show a somewhat lower propensity to contain or support interactions between governance centres than evaluations funded by actors at the EU level or in Germany.

5.5 Conclusion

Across the three broad categories – namely self-organization, context, and interacting governance centres – the empirical evidence presented in this chapter has once again highlighted the internal variation along various dimensions of climate policy evaluations and their ability to facilitate polycentric governance. For example, while formal evaluations are clearly in a majority in the overall database (Chapter 4), they focus only on a small number of policy sectors (e.g., renewables) and omit others (e.g., transport). Turning to the foundational ideas of polycentric governance, formal evaluations are, by definition, not self-organizing in that the original (financial) stimulus to conduct them came from state-actors. However, the 'formal' category is by no means monolithic and presents internal variation that warrants close attention and discussion as this chapter has done. For example, the majority of the formal evaluations that this chapter has analysed were not stimulated by a legal requirement. Second, formal evaluations offered at best a cursory treatment of context—as measured on the various dimensions presented in the first part of this chapter, their ability to disentangle the influence of contextual factors on policy effects is limited. Finally, interactions between governance centres taking place through or enabled by evaluations remains limited among the formal evaluations. It was noticeable that in multiple ways, UK funded evaluations contained fewer interactions and were less-well placed to support interactions than evaluations financed by actors at the EU level or in Germany.

Chapter 6 Informal Evaluation

6.1 Introduction

This chapter contains the results of coding the informal evaluations. The distinction between formal and informal evaluations is based on the evaluation funder (see Chapter 4). Recall that informal evaluations include those funded by 'industry' (including all types of trade associations and lobby groups; 'environmental organizations' such as Greenpeace or Friends of the Earth; non-environmental public interest organizations such as consumer rights organizations; and foundations. Given that the overall database only contained 84 informal evaluations (see Chapter 4), this chapter analyses all of them. The presentation of the results follows the foundational ideas of polycentric governance set out in Chapter 2 namely self-organization of policy evaluation, context in evaluation, and interactions between governance centres.

6.2 Self-organization

Figure 6.1 presents the number of informal evaluations by funder location. It shows that EU level actors funded 40.48% of the informal evaluations, while actors from Germany funded 47.62%, and actors from the UK 11.90% of the informal evaluations. Thus, the contribution of informal evaluations by location of the funder – and thus the self-organizing capacity to evaluate in the polycentric climate governance 'system' in the EU (see Chapter 1) – is palpably uneven across the three governance centres.

Figure 6.1: Funders by location

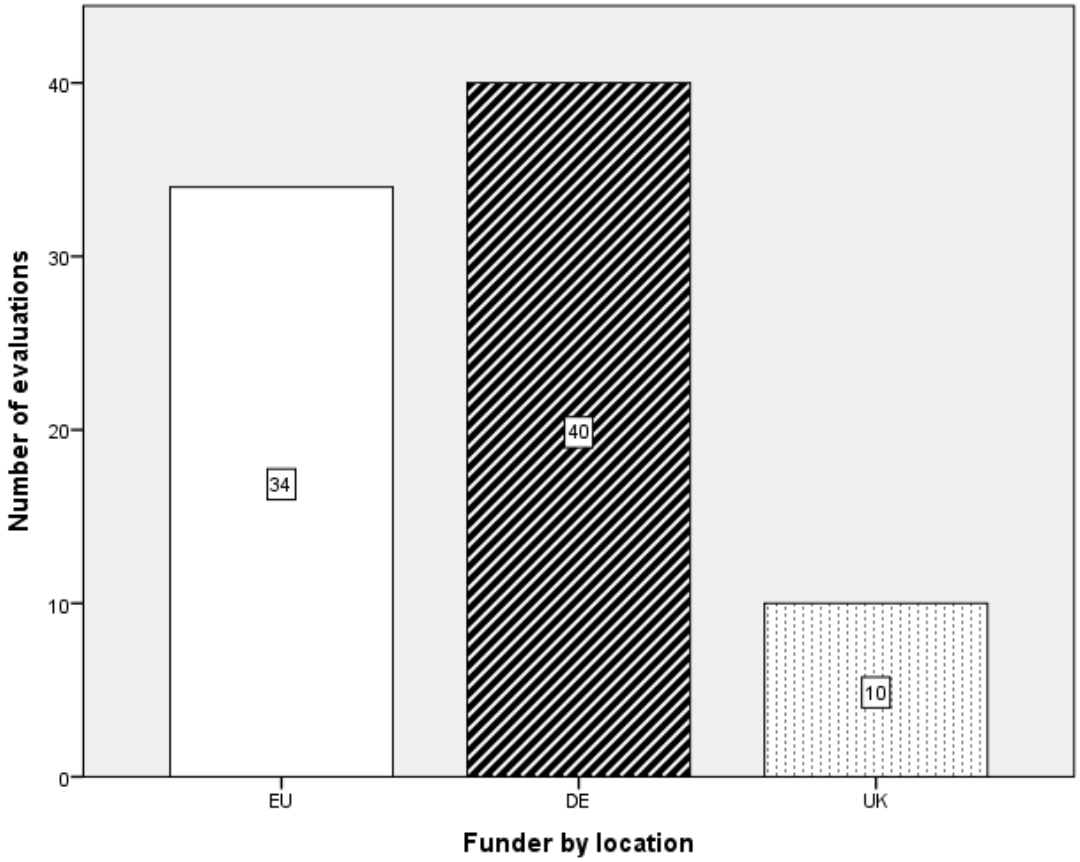
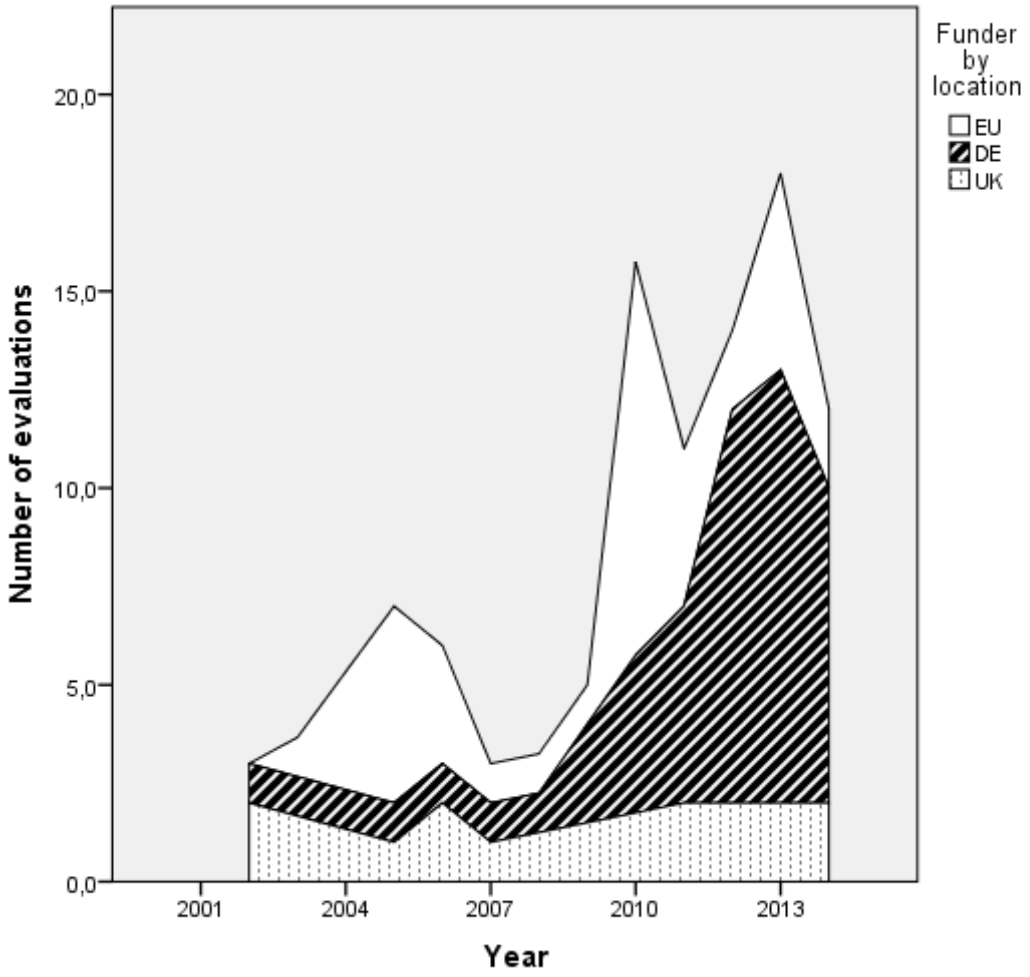


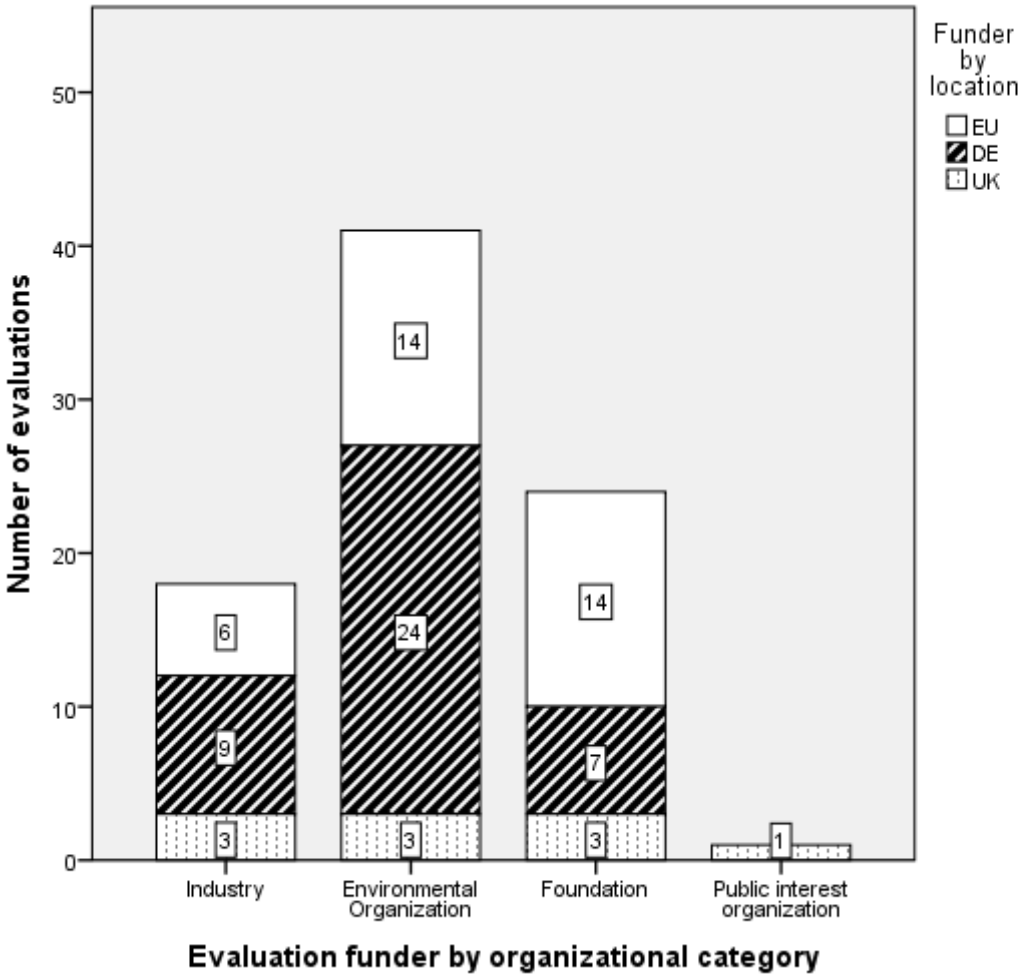
Figure 6.2 in turn demonstrates that informal evaluations first emerged in 2002, and that the number of evaluations funded by actors at EU level, in Germany and in the UK has varied over time. Evaluations funded by EU level actors and actors from Germany mainly drove the notable overall growth in yearly evaluations after 2007, while evaluations financially supported by UK based actors remained at a comparatively low level. While there is an overall growth trend in the number of evaluations over time, it is also clear that the number of informal evaluations tends to fluctuate over time.

Figure 6.2: Informal evaluations by year and funder location



However, the data allow further disaggregating the characteristics of the self-organized evaluations. As Figure 6.3 details, the height of the bars in the chart indicate that environmental organizations funded the largest number of informal evaluations, followed by foundations and, in third place, industry. By contrast, (other) public interest organizations funded a negligible number of evaluations. Evaluation funders from different locations (indicated by the shading of the bars) funded evaluations in most categories, with the exception of public interest organizations, where only UK based funders were active. German environmental organizations such as Greenpeace Germany appeared particularly active in funding evaluations.

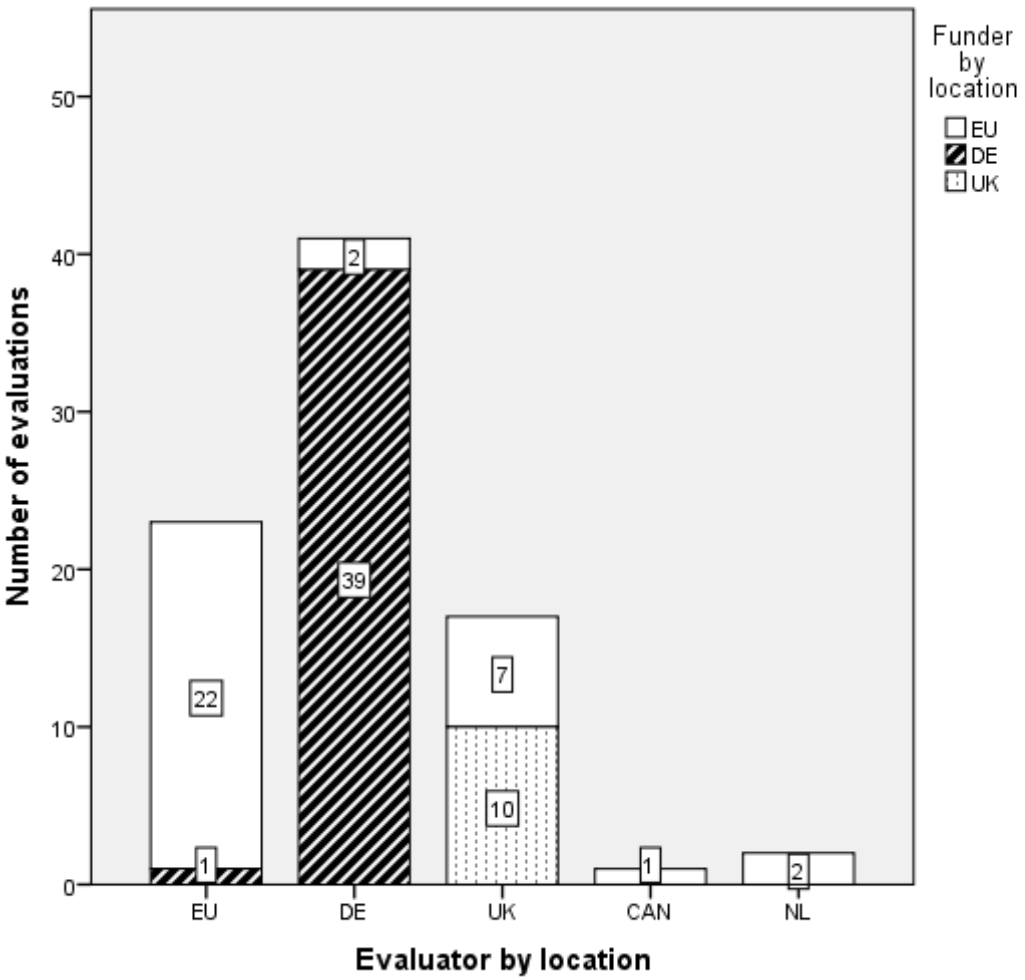
Figure 6.3: Informal evaluation funders by organizational category



Turning to the evaluators, that is, the organizations that conducted the evaluations (note that self-funding is possible), Figure 6.4 demonstrates that the largest number of evaluators are based in Germany (41 evaluations), followed by evaluators at the EU level (23 evaluations) and the UK (17 evaluations). A notable difference with the overall database (see Chapter 4) is that the number of governance centres where the evaluators are located is smaller in that it only includes the EU level, Germany, the UK, as well as Canada and the Netherlands, with evaluators from the latter two governance centres producing only three evaluations in total. A look at the relationship between evaluators and evaluation funders (considering the shading of the individual bars) reveals that, by and large, informal evaluation funders tend to fund evaluators within their own governance centres. In other words, EU level funders mainly funded evaluators at the EU level (that is, evaluators from

multiple EU countries by the definition in Chapter 4), while German funders focused on evaluators from Germany and UK based funders on evaluators from the UK. However, funders at the EU level also funded twelve evaluations conducted by evaluators from other governance centres, namely the UK (7 evaluations), Germany (2 evaluations), the Netherlands (2 evaluations) and Canada (1 evaluation). See Figure 6.4.

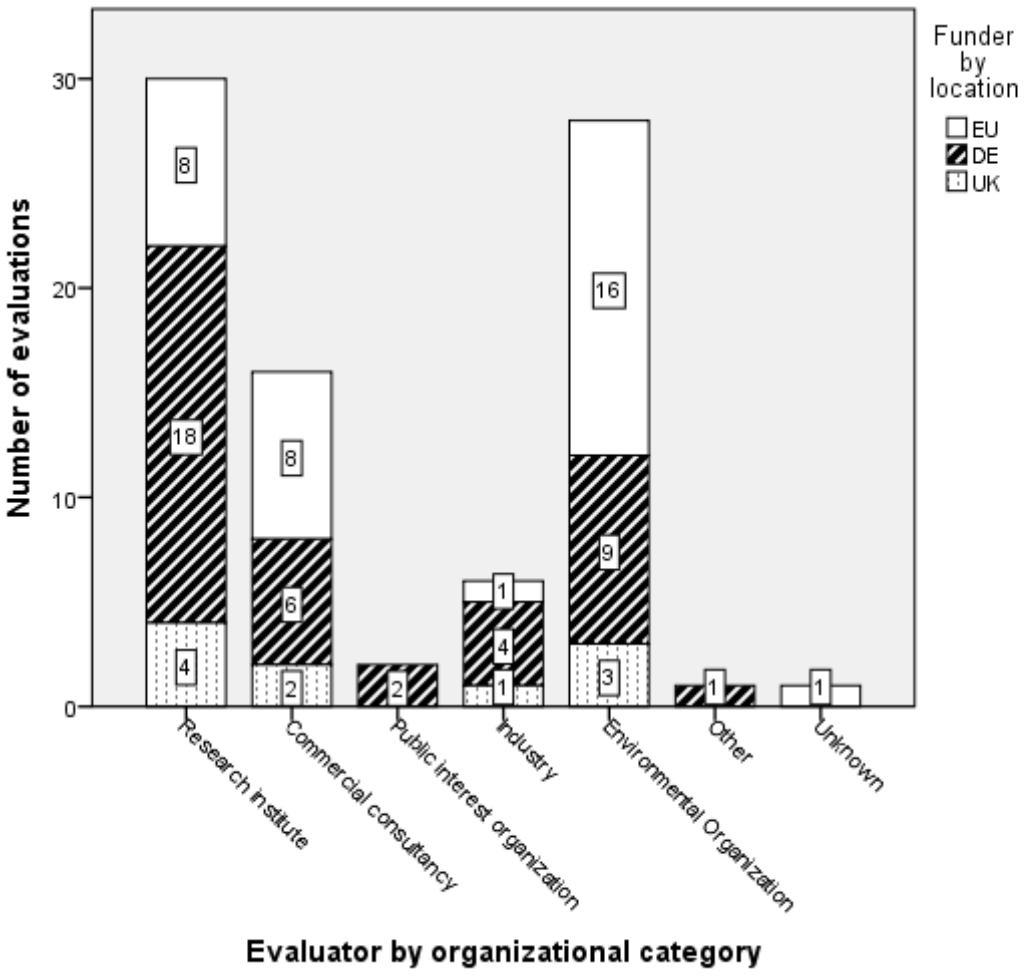
Figure 6.4: Informal evaluations by evaluator and funder location



But what type of organizations did the informal funders choose in order to conduct their evaluations? Figure 6.5 shows that research institutes took a leading role in conducting informal evaluations (30 evaluations), and environmental organizations produced nearly the same amount (28 evaluations). Note that self-funding is a possibility. Importantly, informal actors did not fund any governmental

actors, scrutiny bodies, or courts to conduct evaluations. In contrast to the overall database, commercial consultancies had a lower relative contribution to producing informal evaluations, summing up to 16 evaluations. Informal funders from Germany often funded research institutes to conduct evaluations, while funders at the EU level appeared to have a preference for funding environmental organizations to conduct evaluations.

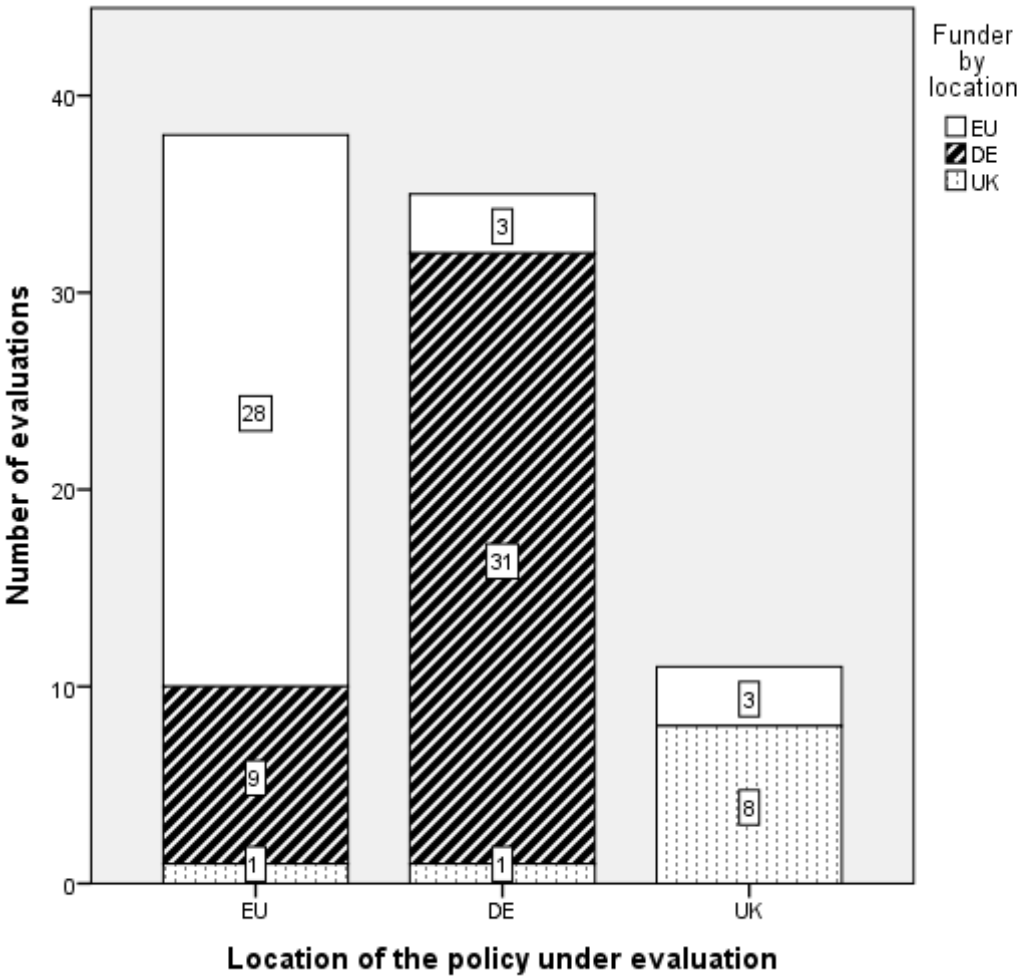
Figure 6.5: Evaluation funders by organizational type



Furthermore, the locus of the policy under evaluation is a somewhat different way to look at these data (see Figure 6.6). Thus, evaluations of climate policy in Germany (national level – 41.67%) and those focusing on the EU (EU level - 45.24%) comprise the lion's share of the sub-set, with evaluations focusing on UK national policy taking a much smaller proportion (13.10%). Taken together, over

85% of all informal evaluations focused on Germany or the EU (see Figure 6.6). Furthermore, looking at the relationship between the location of the evaluation funders and the focus of the evaluation in Figure 6.3 shows that, again, informal funders mainly supported evaluations of policies in their own governance centre. However, in this case, Germany and UK based funders did support a handful of evaluations focusing on EU level climate policy and EU level funders supported nine evaluations focusing on Germany and one on the UK. Altogether, most evaluations are inward-looking, with EU level funders supporting a few evaluations that look beyond its own governance centre and into Germany and the UK.

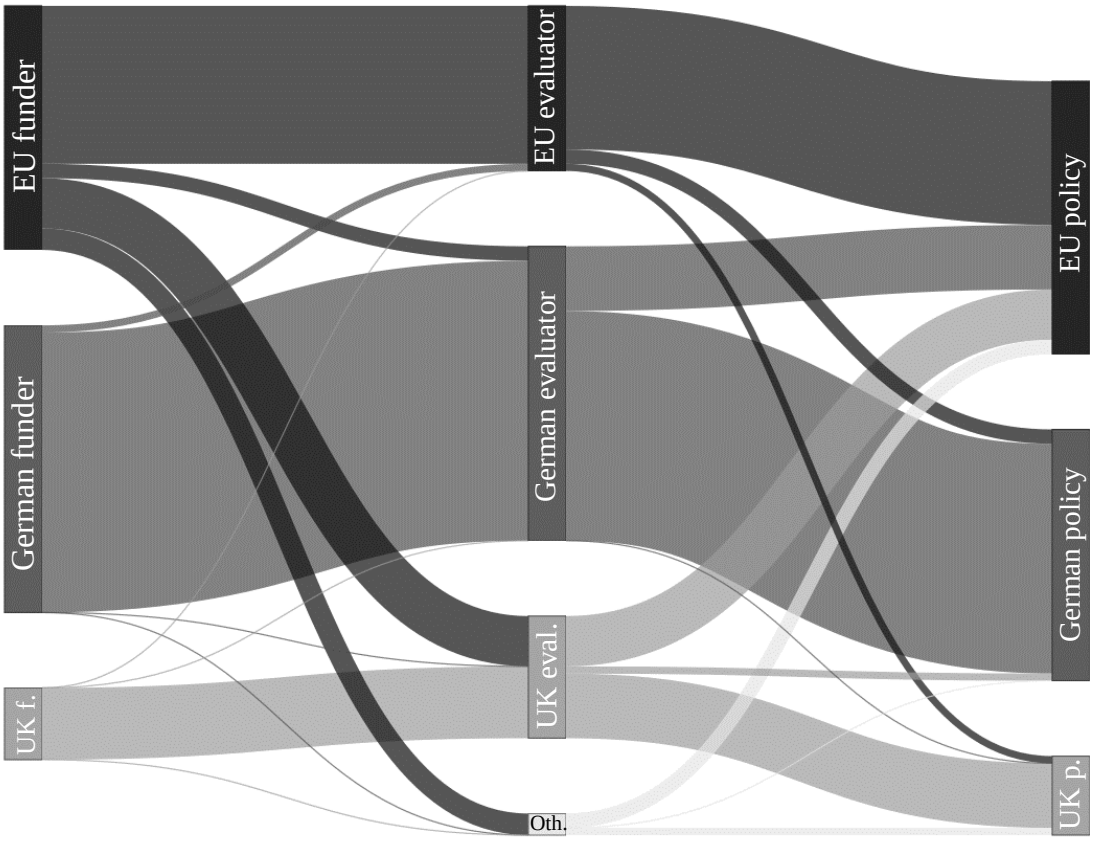
Figure 6.6: Informal evaluations by location of policy under evaluation



Analogous to Chapters 4 and 5, Figure 6.7 draws together the information on the location of the evaluation funders, the evaluators, as well as the location of the

policy under evaluation into a single Sankey figure. The thickness of the connectors between funders, evaluators and the policy represents the number of evaluations with the respective characteristics (i.e. a thicker line means that there are more evaluations with these characteristics). A readily visible feature is that German funders mainly funded German evaluators in order to conduct evaluations of German climate policy, with similar patterns for the EU level and for the UK. However, we can also observe that EU level actors funded a sizeable number of UK based evaluators to evaluate EU level climate policy; similarly, German funders funded a number of German evaluators in order to assess EU level climate policy.

Figure 6.7: Location of evaluation funders, evaluators and policy under evaluation

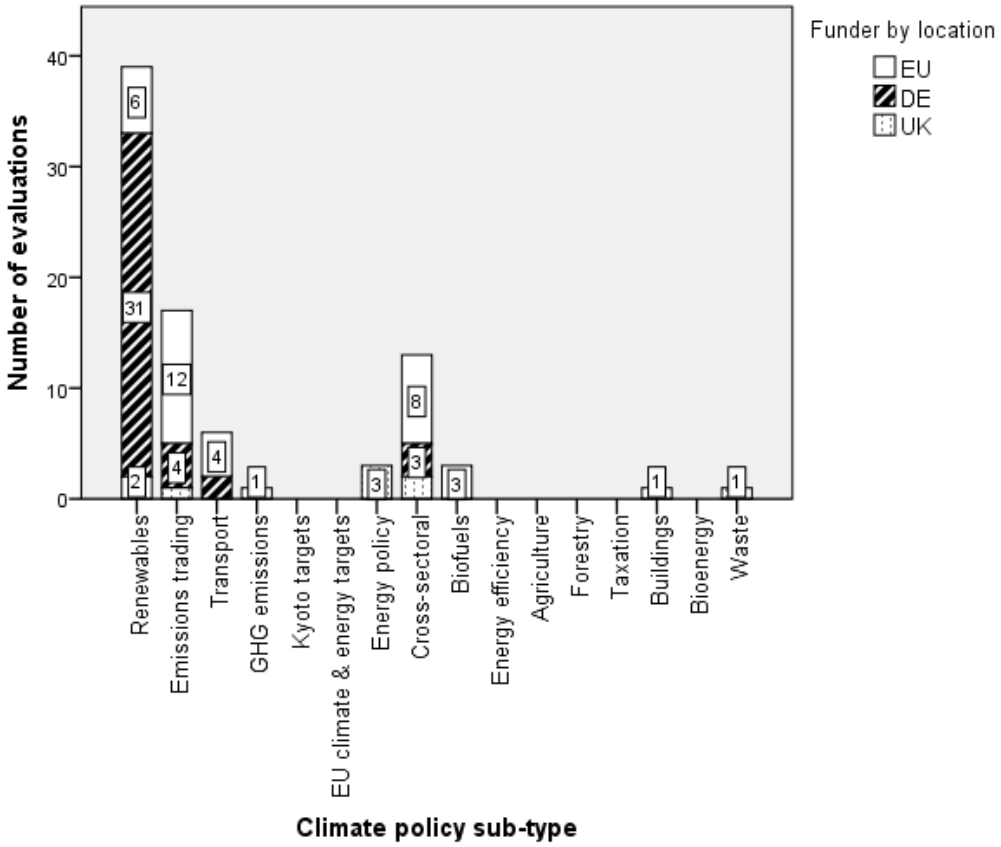


Note: the thickness of the connectors represents the number of evaluations with the respective characteristics.

An additional way to look at the informal evaluations is to consider the policy sub-type on which the evaluations focus. Figure 6.8 summarizes the relevant data. Similar to the overall database (see Chapter 4), by far the greatest number of informal evaluations focus on renewables policy, followed in considerably smaller

numbers by emissions trading and cross-sectoral policy (which includes evaluations that focus on more than one of the sub-sectors). Considering the location of the evaluation funders shows that Germany based evaluation funders mainly supported evaluations focusing on renewables, and EU level evaluation funders with a stronger focus on emissions trading, cross-sectoral policy, but also renewables. UK based evaluation funders were the only ones to support evaluations of more general energy policy, such as emissions from the entire power sector.

Figure 6.8: Informal evaluations by climate policy sub-type



However, there are also additional ways of further unpacking the concept of self-organization. Another variable to consider is the extent to which a legal requirement drives informal actors to evaluate. Unsurprisingly, but nevertheless importantly, none of the evaluations included in the informal cohort responded to a legal requirement.

Another way of looking at self-organization is to consider the extent to which the informal evaluations form parts of larger, continuous evaluation efforts or whether they are relatively ad-hoc. Figure 6.9 reveals that the vast majority of informal evaluations (72.62%) are ad-hoc. Only 27.38% comprise evaluation efforts that form part of larger cycles, evidenced by reference to prior or subsequent evaluation activities. Informal evaluation funders are thus generally either unwilling or unable to maintain ongoing evaluation exercises. Actors based in the UK funded no continuous evaluations, unlike funders at EU level and in Germany, which respectively funded 9 and 14 continuous evaluations.

Figure 6.9: Temporal nature of evaluations by funder location

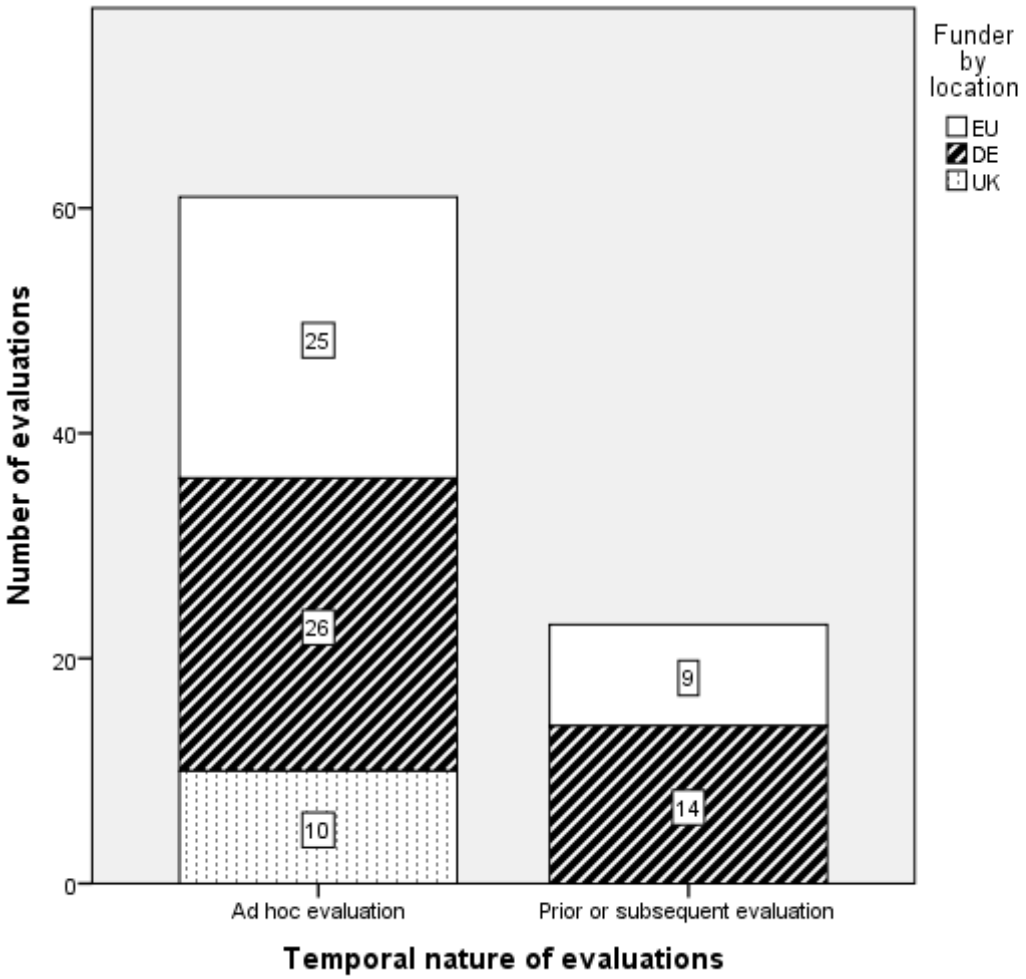
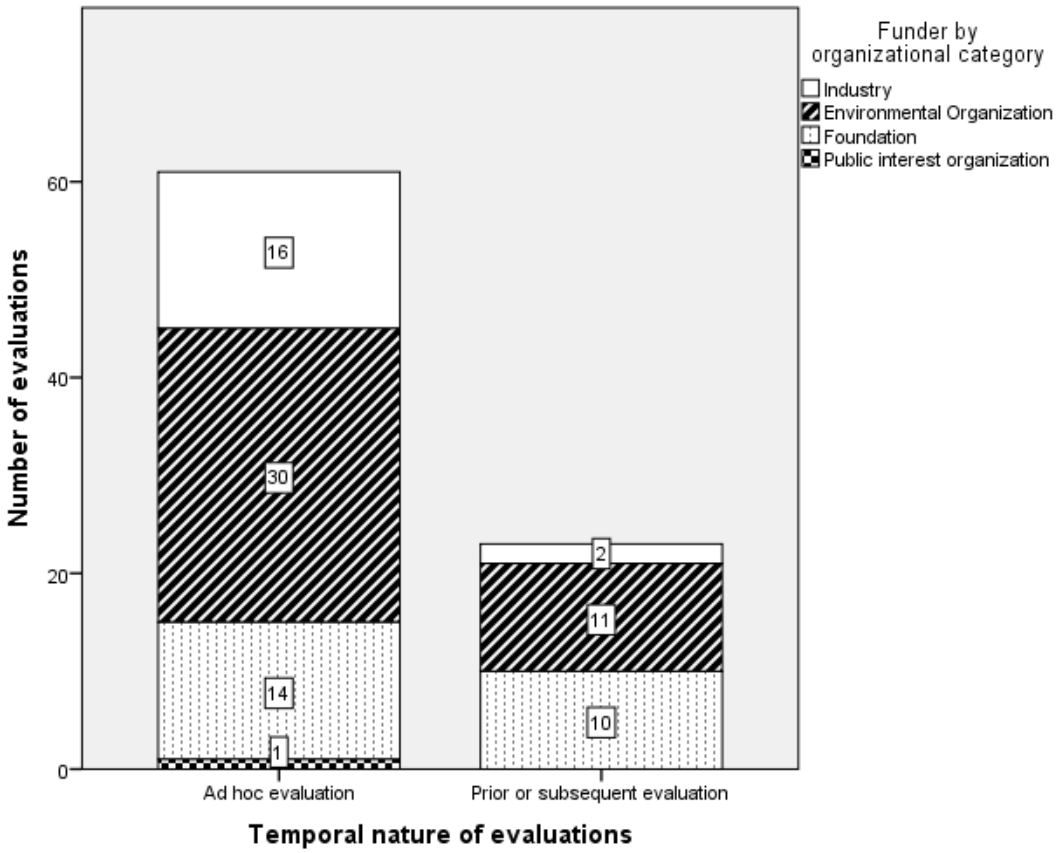


Figure 6.10, in turn, disaggregates the continuous versus ad-hoc evaluation variable by funder category. It reveals that environmental organizations funded both

the largest number of ad-hoc and continuous evaluations. Industrial actors funded a very small number of continuous evaluations, but foundations funded a comparatively large amount of continuous evaluations (41.67% of the evaluations funded by foundations in total).

Figure 6.10: Temporal nature of evaluations by funder category



- In sum, this section reveals that, there are important patterns in who funds and who conducts informal evaluations, and what they evaluate, including over what time scale.
- Given that the number of informal evaluations in the overall database only comprises 84 evaluations, it is clear that there are limits to self-organizing (i.e. non-state) climate policy evaluation in the EU across the three governance centres considered here. By the definition used in this thesis (see Chapter 4), the vast majority of evaluations were not self-organizing.

- Looking across the findings, the most striking point is that in many ways, there is a strong congruence between evaluation funders, evaluators and the focus of the evaluations, which tend to focus on the same governance centres (i.e. informal funders from Germany tend to fund German evaluators focusing on German climate policy and likewise for informal funders from the UK). EU level evaluation funders are somewhat different in that while they also focus mainly on their own governance centre, they also fund evaluators in other countries, and some of their evaluations focus on Germany and the UK. Especially in Germany, environmental organizations often take a lead in funding and conducting informal evaluations.

6.3 Context

This section addresses the contextual dimension of climate policy evaluation. Figure 6.11 summarizes the group of contextual variables that are included in a context index (see below). Chart A in Figure 6.11 addresses time horizons in evaluation, namely, as Chapter 4 explained, the number of years which an evaluation considers. For example, this could include how far the evaluation looks back in a historical review of the policy, or to what extent it presents data over a longer time span. Chart A reveals that informal evaluations most frequently consider time spans of over twenty years. The remainder of the evaluations scatter more or less equally among the other categories. Looking at the evaluation funders (the shading of the bars) shows that informal funders from all three governance centres (EU level, Germany and the UK) financially supported evaluations in each category, and Germany and the UK showed a slightly higher propensity to fund evaluations with a very long time span from 16 to 20 years and more.

Figure 6.11: Contextual variables in informal evaluations

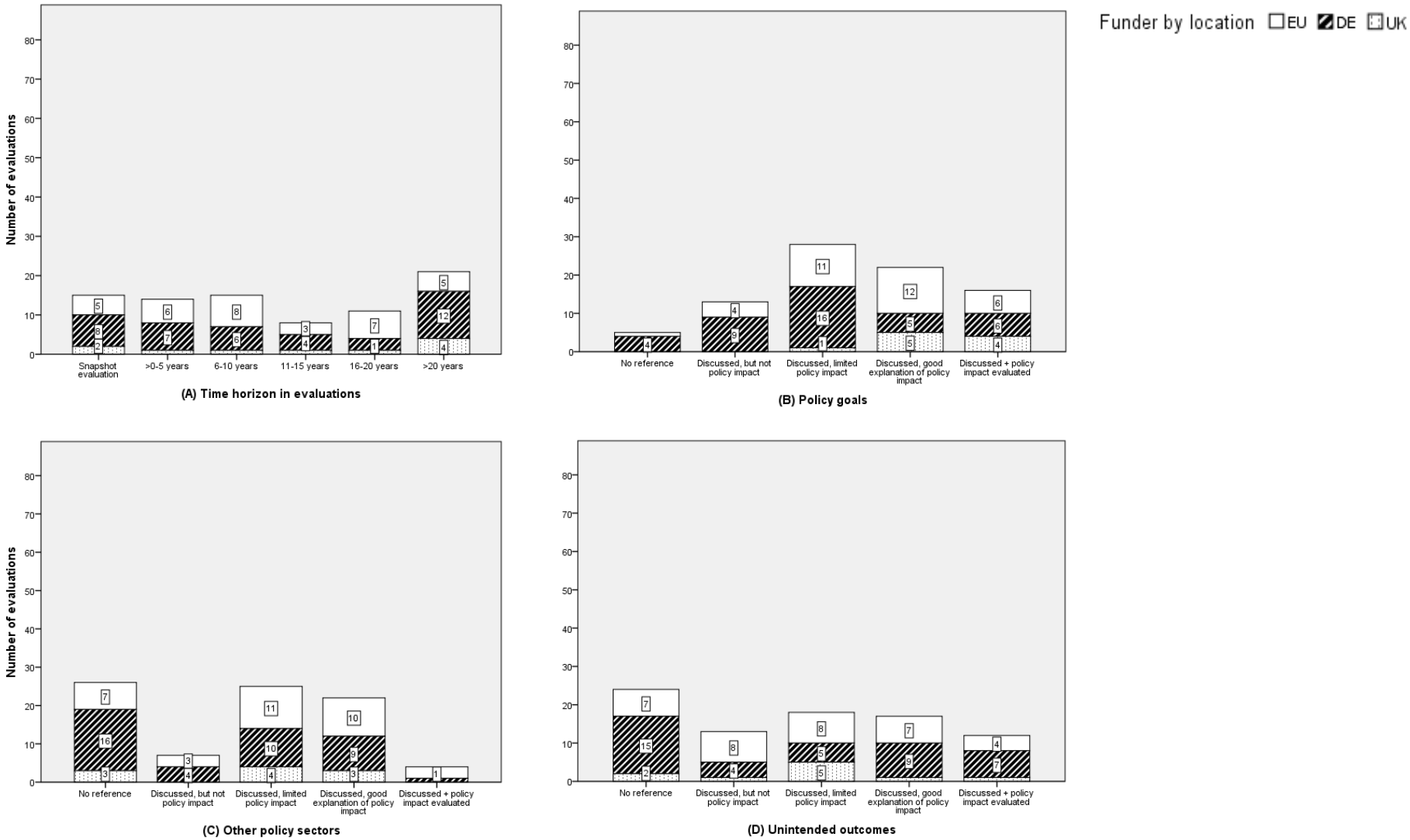
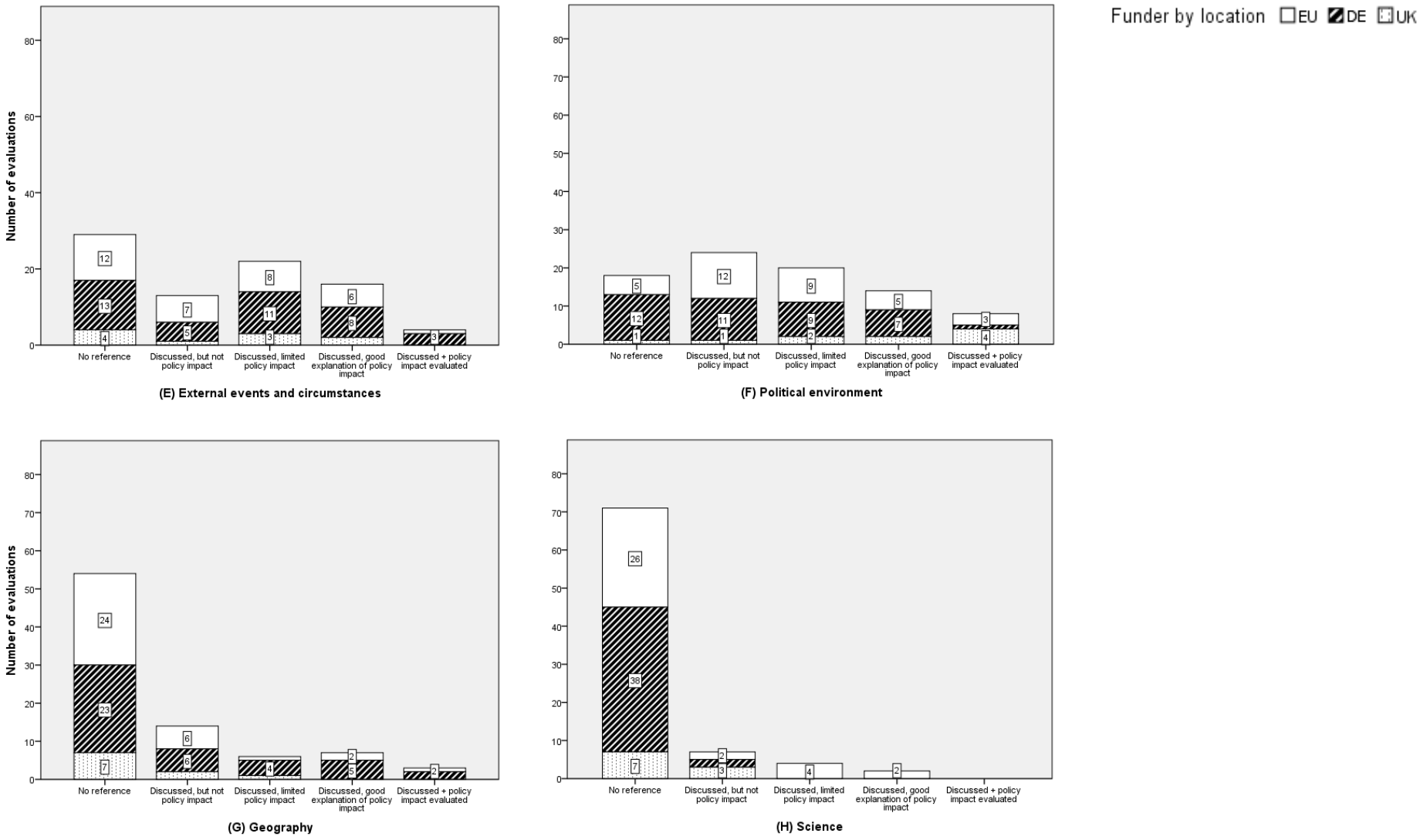


Figure 6.11 (continued)



The second contextual variable considered in Figure 6.11 is the extent to which the evaluations make reference to extant policy goals (e.g. the greenhouse gas reductions targets at the national or at the EU level).[37] Chart B in Figure 6.11 demonstrates that most informal evaluations discussed policy goals to some degree, but less than half of the evaluations (45.24%) considered policy goals well (score = 3) or extensively (score = 4). Looking at the distribution of the evaluation funders (the shading of the bars) highlights that evaluations funded by actors based at the EU level and in the UK showed a somewhat higher propensity to engage more deeply with policy goals than evaluations funded by actors from Germany.

The third contextual category relates to references to other policy sectors with a view to linkages with the policy at the centre of the evaluation. For example, this may relate to the extent to which a policy to support wind power may link with spatial planning and nature protection policies. Chart C in Figure 6.11 demonstrates that about a third of the informal evaluations made no reference whatsoever to linkages with other policies. The proportion of evaluations that discussed this dimension and then evaluated the impact of the interactions on policy outcomes remains small with just under five evaluations. Thus, in general, by the evidence from the evaluations, this contextual dimension received relatively little attention. The distribution of evaluations by funder is relatively even across the governance centres. However, notably, no evaluations funded by actors based in the UK received a (maximum) score of 4.

Chart D in Figure 6.11 presents data on the extent to which the informal evaluations considered unintended policy outcomes, because these outcomes frequently emerge from interactions between a policy and its context (see Chapter 2). In other words, unintended and especially unforeseen effects may emerge precisely because contextual circumstances interact with the policy in unexpected ways. The

[37] Recall that this variable and all the remaining variables in Figure 6.11 were scored on a 0-4 scale, where 0 = no reference to dimension; 1 = dimension discussed, but no explanation of how this dimension impacts policy outcomes; 2 = dimension discussed, but limited explanation of how this dimension impacts policy outcomes; 3 = dimension discussed, and good explanation of how this dimension impacts policy outcomes; 4 = dimension discussed and impact on policy outcomes evaluated extensively (for further details, see Chapter 4 and Appendix 3).

bar chart demonstrates that 27.38% of the informally-funded evaluations paid no attention whatsoever to unintended side effects. A few evaluations briefly addressed side effects, but only 34.52% of the evaluations paid good or extensive attention to side effects and elaborated on related policy impacts. The distribution of evaluations by location of the funder is relatively proportionate, even though it should be noted that UK based informal evaluation funders financially supported only two evaluations that engaged well or extensively with unintended policy outcomes.

One of the contextual factors that may have considerable impact on policy outcomes concerns external events and circumstances —think, for example, of the global financial crisis that began in September 2008, marking the onset of the global recession, which has been cited as one of the reasons why the EU may easily reach its 2020 greenhouse gas reduction targets (Jacobs, 2012). Chart E in Figure 6.11 reveals that, similar to the previous category, 54.52% of the informal evaluations did not address external events and circumstances at all. Evaluations in the two highest-scoring categories (i.e. scores of 3 or 4) represent 28.57% of the overall evaluations funded by informal actors. Turning to the distribution of evaluation funders (the different shading of the bars) shows a relatively even distribution, but again, informal actors based in the UK did not fund any evaluations that received the highest score (4) on this criterion.

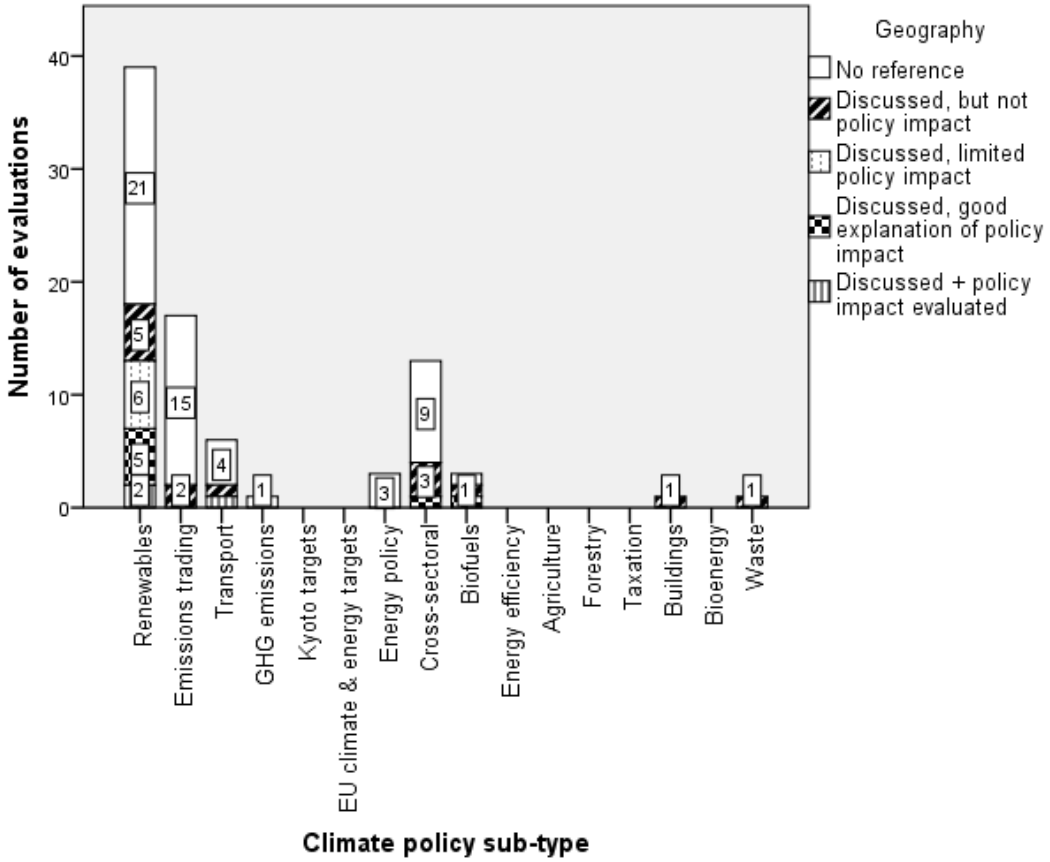
Furthermore, Chart F in Figure 6.11 shows data on the extent to which evaluations focused on the broader political environment in which a climate policy is placed, including institutions, political actors, and events. In contrast to the earlier dimensions, the category with the greatest number of evaluations is that which mentions or briefly discusses the political environment, but does not deal with its policy impact. The number of evaluations that discuss the political environment in depth and evaluate its policy impact remains relatively low (25%). The distribution by location of the evaluation funders (shading of the bars) is relatively proportionate.

The next contextual variable concerns (physical) geography, because not all policy approaches will necessarily work across all geographical contexts, especially in the case of climate policies that have a clear geographical dimension (such as the siting of wind turbines, agriculture, or biomass and forestry). Chart G in Figure 6.11 details the extent to which evaluations took account of this dimension. In general, the

level of attention to geography across all evaluations is low, with 64.92% of the evaluations making no reference to this dimension. Only 11.90% of the evaluations focused on the policy-impact of the geographical context. The distribution by evaluation funders (bar shading) shows that evaluation funders based in Germany funded a disproportionately large number of evaluations that paid deeper attention to geography; UK based funders financially supported no evaluations that discuss the impact of geography on climate policies in depth (i.e. a score of 4).

But geography may matter more with some policies than with others; for example, tidal energy depends on the availability of a coastline with certain characteristics, whereas emission limits on a traded product such as cars may be less dependent on geographical factors. Therefore, Figure 6.12 details attention to the assessment of geography in evaluations by policy type. In light of the overall distribution of these data, it is perhaps not surprising that renewables policy features at high levels in all categories on the familiar scale; however, it is also notable that when evaluations do focus on geography and its impact on policy outcomes, it tends to be in the area of renewable energy, with much lower attention to the role of geography in influencing climate policy outcomes in other policy areas. But even with renewables, only 18 evaluations evaluated the impact of geography in depth (i.e. they received a score of 4 on the relevant scale).

Figure 6.12: Attention to geography in evaluations by climate policy sub-type

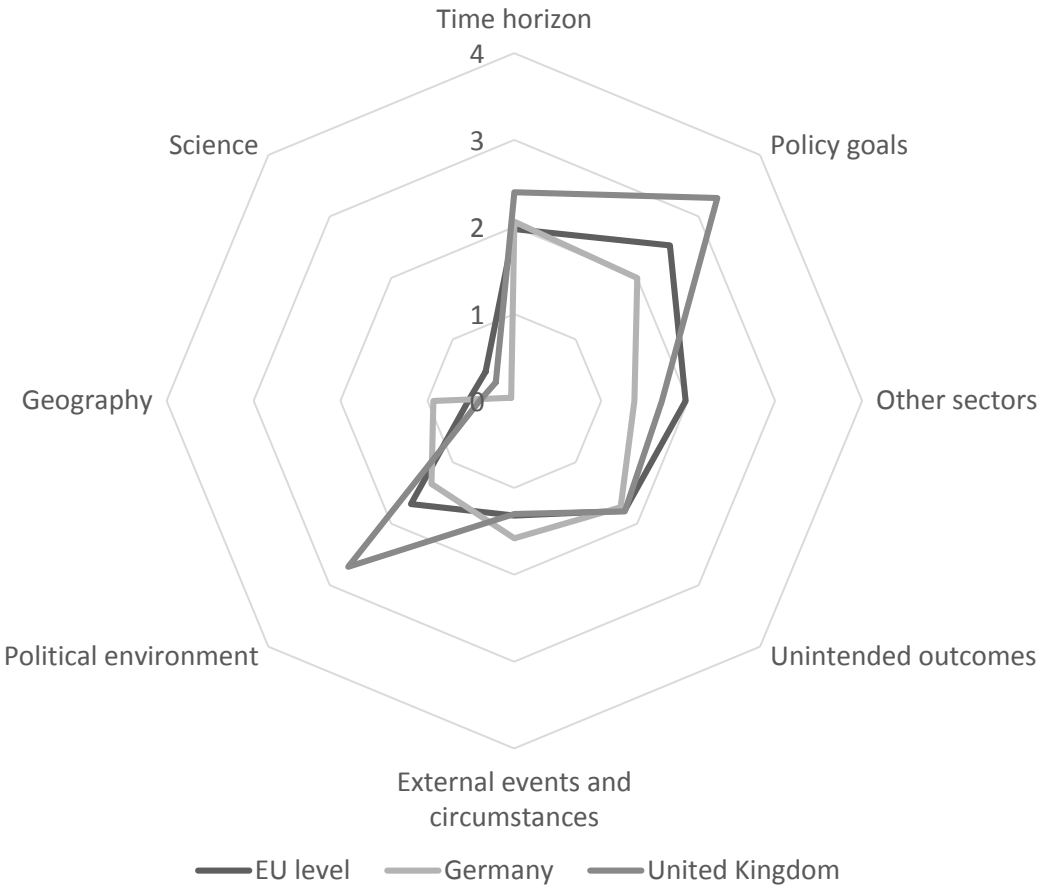


Turning back to the discussion of Figure 6.11, Chart H presents data on the extent to which climate policy evaluations dealt with a discussion of (climate) science. In the area of climate change policy, the science of climate change provides the indispensable backdrop against which climate policies are formulated. However, the vast majority of evaluations (82.14%) made no reference to climate science, so if at all, this dimension was considered rather implicitly. While five evaluations (5.95%) made general references to climate science, only six contained more detailed explorations. Noticeably, informal actors at the EU level funded all of the small number of evaluations (6 or 7.14%) that addressed climate science in greater detail. The latter included for example discussing findings from the IPCC Assessment Reports and the potential consequences of unabated climate change.

In order to visualize the eight contextual variables discussed above together, Figure 6.13 presents their averages in a spider diagram. In line with the procedure

described in Chapter 5, the time variable was transformed onto a 0-4 point scale.[38] On the diagram, the vertical numbers represent the familiar scale (0-4, see above), and the individual contextual variables feature on each ray of the diagram. Considering the data by governance centre (the differently shaded circles in the middle of the diagram) demonstrates that evaluations funded by informal actors from the UK were particularly strong in contextualizing with a view to the political environment, as well as policy goals. Informal evaluations supported by actors at the EU level were stronger than the others in considering other sectors.

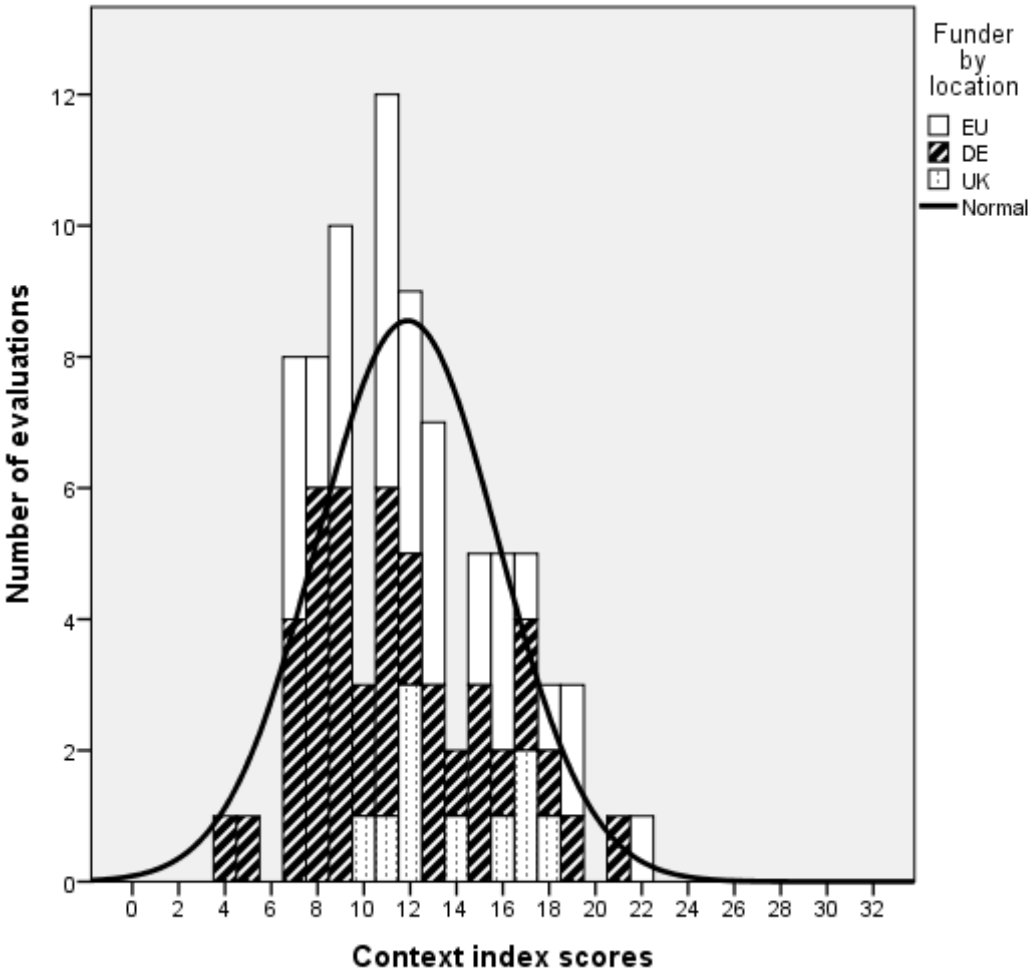
Figure 6.13: Average scores on contextual variables by governance centre



[38] I conducted the norming subtracting 1 from each time horizon score, and multiplying the result by 4/5.

Following the exploration of these individual contextual variables, I calculated Pearson correlations in order to detect any possible relationships between the variables (see Table A4.2 in Annex 4). By and large, the correlations among the contextual variables from Figure 6.11 are weak and often statistically insignificant. Therefore, in order to construct an overall index of attention to context, I calculated the sum of the contextual variables per evaluation in order to reflect the fact that an evaluation may be strong on one dimension, but not necessarily on the others (I transformed the time variable onto a 0-4 scale, see above). The distribution's average is $M = 11.89$, with a standard deviation of $SD = 3.92$. The minimum and maximum scores are 4 and 22, respectively. This range is compared against the theoretical minimum (0) and the theoretical maximum (32), showing that the distribution in covers a little more than half, but certainly not all, of the theoretical range. The number of evaluations with each index score were then plotted onto a histogram, which shows a near-normal distribution (see Figure 6.14).

Figure 6.14: Index of contextual variables in informal evaluations

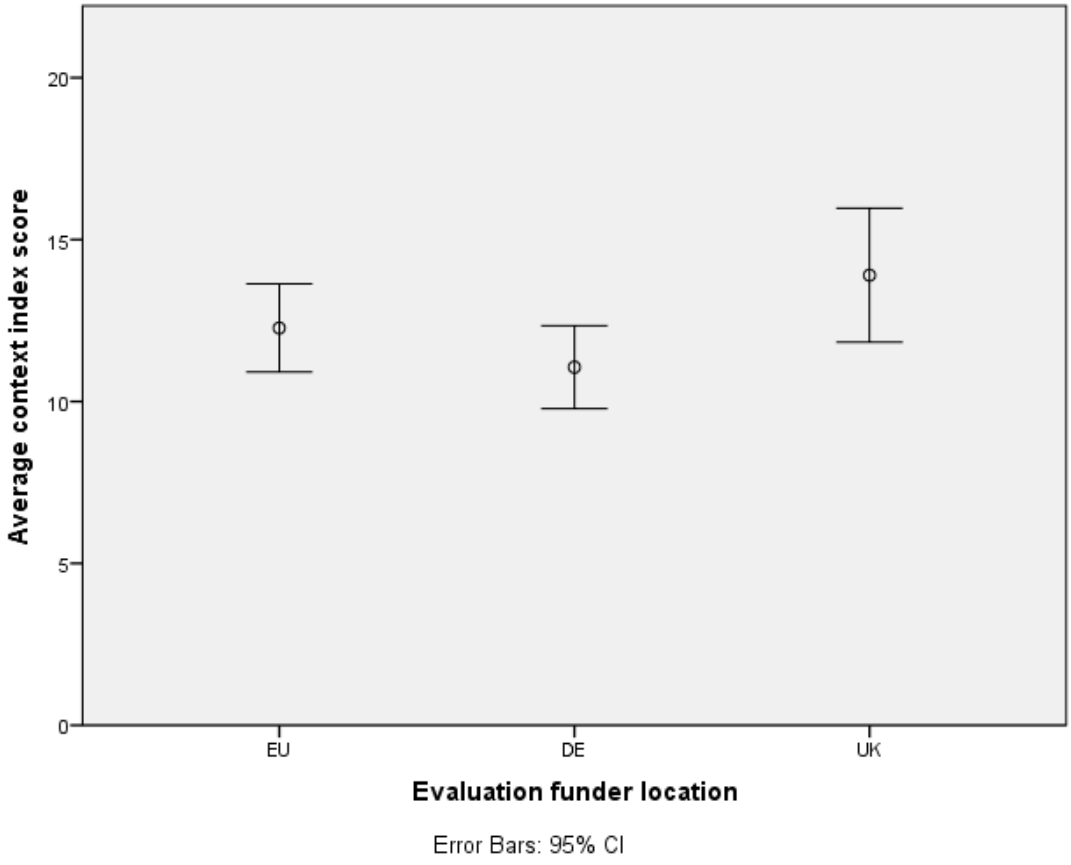


Evaluation funders from the three governance centres (the EU level, Germany and the UK) make somewhat different contributions to the overall distribution. Evaluations funded by actors in Germany cluster at the lower end of the spectrum with $M = 11.06$ ($SD = 4.00$), followed by the evaluations funded by actors at the EU level ($M = 12.27$, $SD = 3.90$) and more towards the higher end of the spectrum, from the UK ($M = 13.90$, $SD = 2.89$).[39] Figure 6.15 presents these averages graphically, demonstrating some overlap in the confidence intervals. A one-way Analysis of Variance (ANOVA) in order to compare these differences statistically revealed a marginally significant result with $F(2, 81) = 2.46$, $p = .092$. In other words, the

[39] Given that the index score includes some ordinal variables, the means and means testing are indicative only.

source of evaluation funding appears to hang together with extent to which the evaluation considers contextual elements; attention to context in policy evaluations evidently differs across different governance centres.

Figure 6.15: Average scores on contextual index by governance centre



As Chapters 2 and 5 have already argued, in addition to these directly measurable contextual variables, there are also other, more indirect characteristics of evaluations that pay attention to context. The first of these relates to evaluation methods, and especially their number, as a greater number of evaluation methods points to an effort to try to capture as many policy effects as possibly by means of triangulation. The idea of triangulation comes from the notion that each method has unique strengths and weaknesses – combining several methods can thus help to capture multiple policy effects, including those related to context. Figure 6.16 details that by far the most widely used evaluation method is a document analysis (used by

86.60% of the informal evaluations), followed by questionnaires/interviews (used in 22.62% of the evaluations) and modelling (used in 22.62% of the evaluations). Informal evaluations used other methods, such as cost-benefit analysis, expert input, or stakeholder involvement, much less. Turning to the contribution of evaluations by funder location (shading of the bars in Figure 6.16) indicates a relatively proportionate distribution, although evaluation funders in the UK supported fewer evaluations that used modelling than those at the EU level or in Germany.

Figure 6.16: Types of methods in informal evaluations[40]

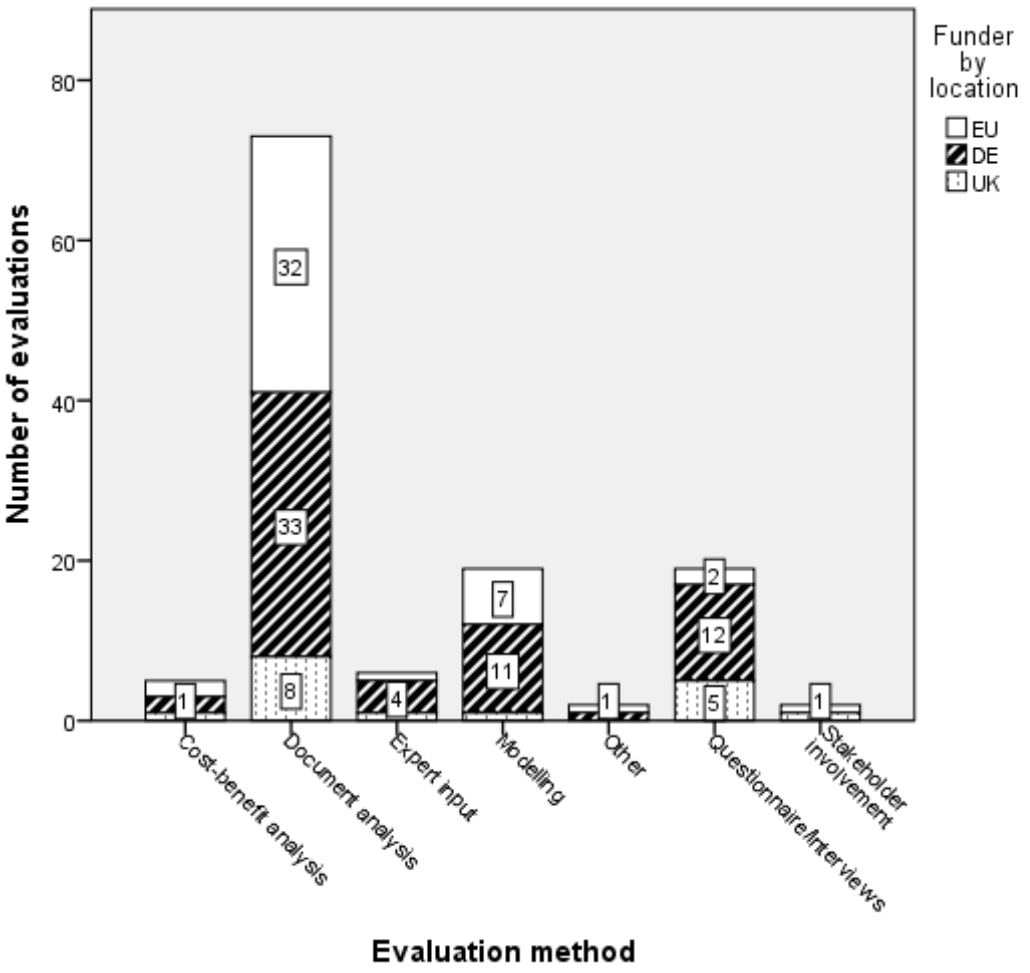
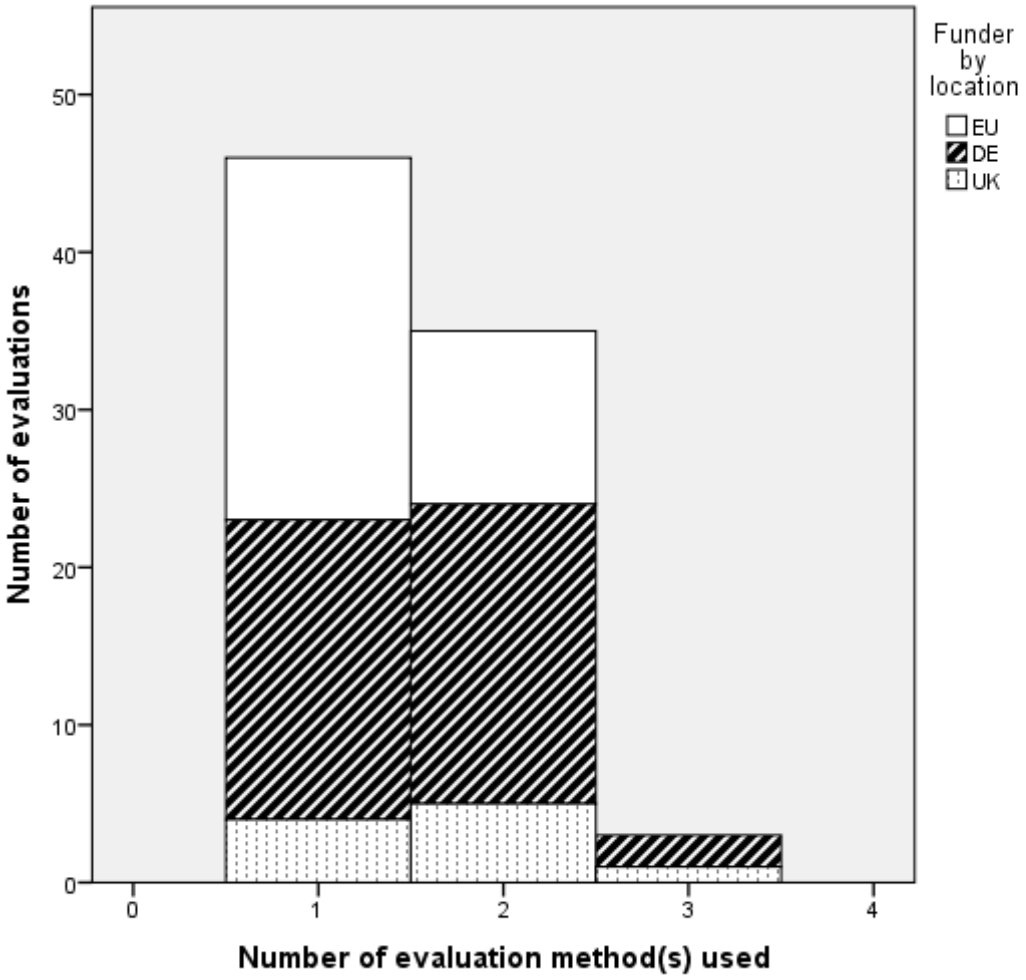


Figure 6.17 in turn presents the number of methods reported in the informal evaluations. Strikingly, the highest number of methods used in any informal

[40] Multiple responses possible.

evaluation is three, while about half of the evaluations used only a single method. A little more than a third of the evaluations used two methods. Overall, the number of methods in informal evaluations remains at the lower end of the spectrum. Looking at the distribution by evaluation funder location shows a relatively even distribution; calculating averages confirms this notion, where evaluations funded by actors at the EU level used on average *M* = 1.32 methods (*SD* = .48), evaluations funded by Germany based actors returned on average *M* = 1.58 methods (*SD* = .59) and the UK *M* = 1.70 (*SD* = .68).

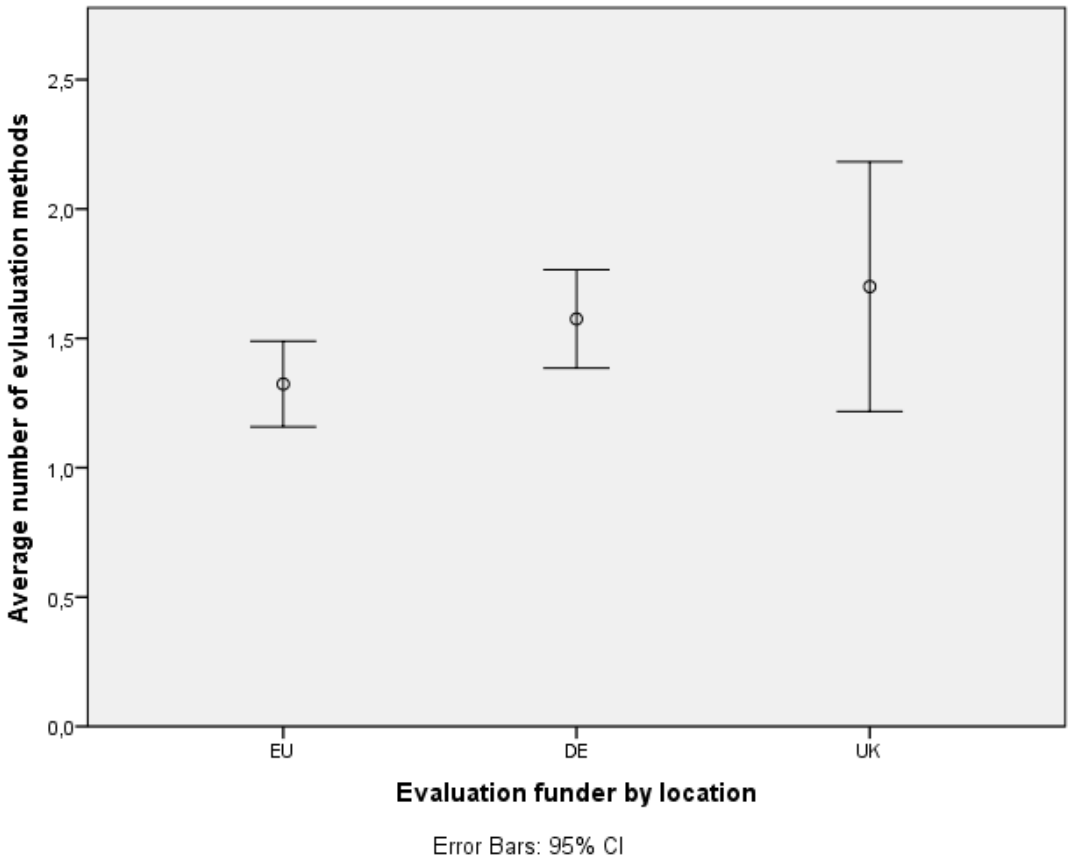
Figure 6.17: Number of methods in informal evaluations[41]



[41] Multiple responses possible.

Figure 6.18 presents these averages graphically, demonstrating the higher level of methods in informal evaluations funded by UK based actors, but also showing that the confidence interval is much larger, which indicates a greater spread in the number of methods used in these evaluations. An ANOVA to statistically compare these averages revealed marginally significant results with $F(2, 81) = 2.68$, $p = .075$.

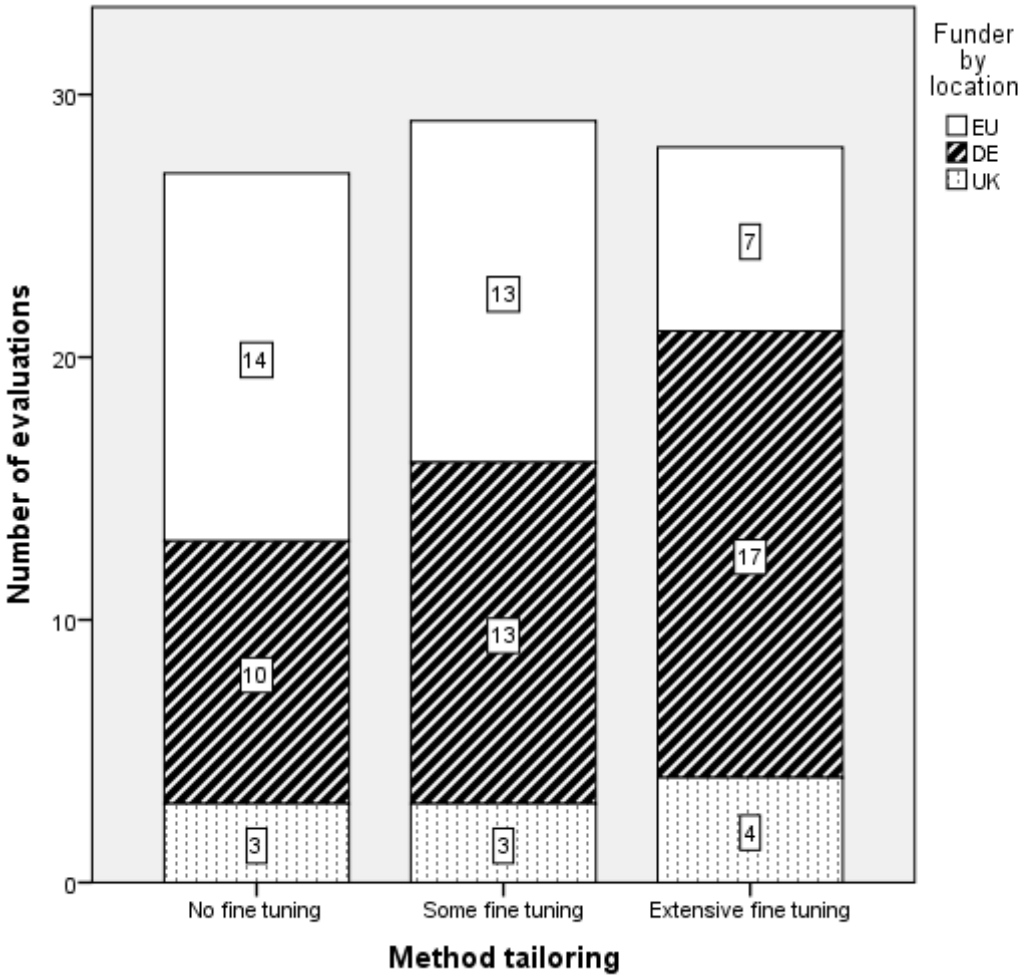
Figure 6.18: Average number of methods in informal evaluations by funder location



In addition to the number of evaluation methods, another important way in which an evaluation may recognize context is through the extent to which the evaluator tailors their methods to the policy in question. An example may be calibrating a model or specifically designing or adjusting a questionnaire for the evaluation. The respective data reported in Figure 6.19 were extracted from the methodological description in each informal evaluation. The informal evaluations

separate into almost three equal groups on this criterion. In other words, a third of the evaluations exhibited no tailoring of the method at all (i.e. by and large they used 'off the shelf methods'), while another third used some tailoring/calibrating, and the final third used extensive fine tuning, which may also include developing a new instrument, such as a questionnaire or a model, for the evaluation. The distribution by evaluation funders is again fairly even, although funders based in Germany funded a greater number of evaluation with extensive tailoring and funders at EU level financially supported comparatively fewer studies that did so.

Figure 6.19: Methodological 'tailoring' in informal evaluations

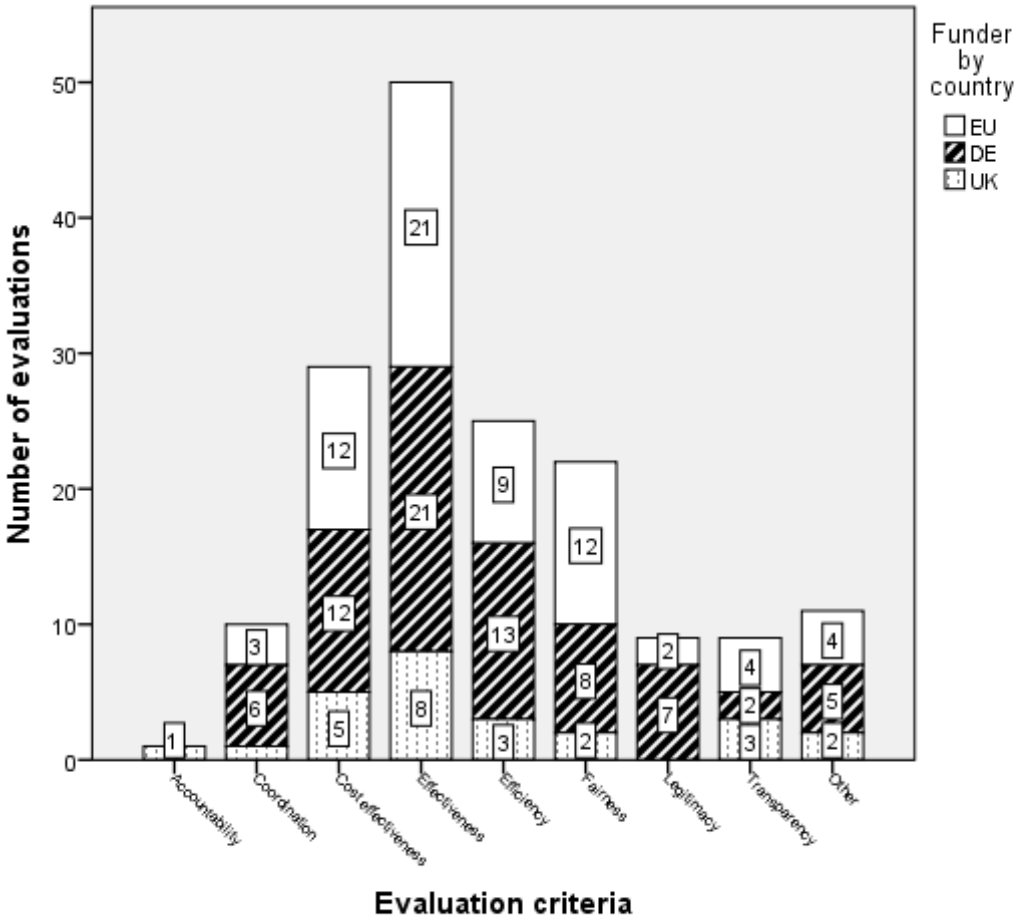


In addition to the methods that informal evaluations use, another indicator of attention to contextual effects comprises the evaluation criteria. Analogous to the argument on methods, a diversity of criteria suggests that the evaluators endeavour to

detect contextual effects through multiple perspectives expressed by a variety of criteria. Figure 6.20 describes the criteria that the informal evaluations used.[42] A strong majority of the informal evaluations (59.52%) used the criterion of effectiveness, which relates to the extent to which a policy reaches its stated or implied aims. This is followed by cost efficiency, overall efficiency, and fairness. On the whole, informal evaluations subjected climate policy to a range of criteria. The distribution by evaluation funder was relatively proportionate, although it should be noted that actors in the UK did not fund any evaluations that challenged the legitimacy of the climate policy in question.

[42] Multiple mentions possible.

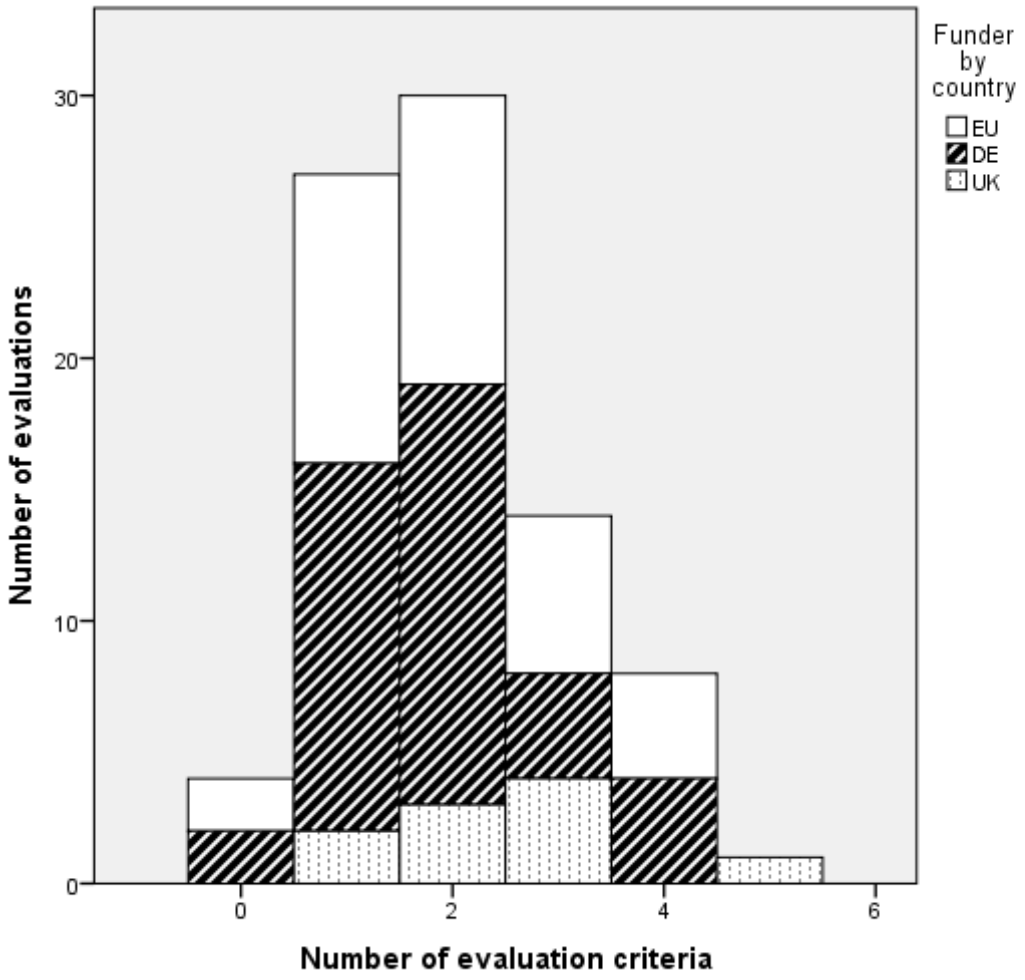
Figure 6.20: Types of criteria used in informal evaluations[43]



How many criteria did the informal evaluations use overall? Figure 6.21 plots the number of evaluations against the number of criteria they used. The first two bars indicate that the greatest number of evaluations either assessed the climate policy against one or two criteria. A smaller number of informal evaluations used three or four criteria, while only five evaluations used five criteria. On average, informal evaluations used $M = 1.98$ criteria ($SD = 1.09$). The distribution by evaluation funder is by and large proportional, which also becomes clear by looking at how many criteria the informal evaluations from each governance centre used, with the EU level ($M = 1.97$, $SD = 1.11$), Germany ($M = 1.85$, $SD = 1.03$), and the UK ($M = 2.50$, $SD = 1.18$) relatively close together. A one-way ANOVA to compare these averages yielded statistically insignificant results, $F(2, 81) = 1.45$, *ns*.

[43] Multiple responses possible.

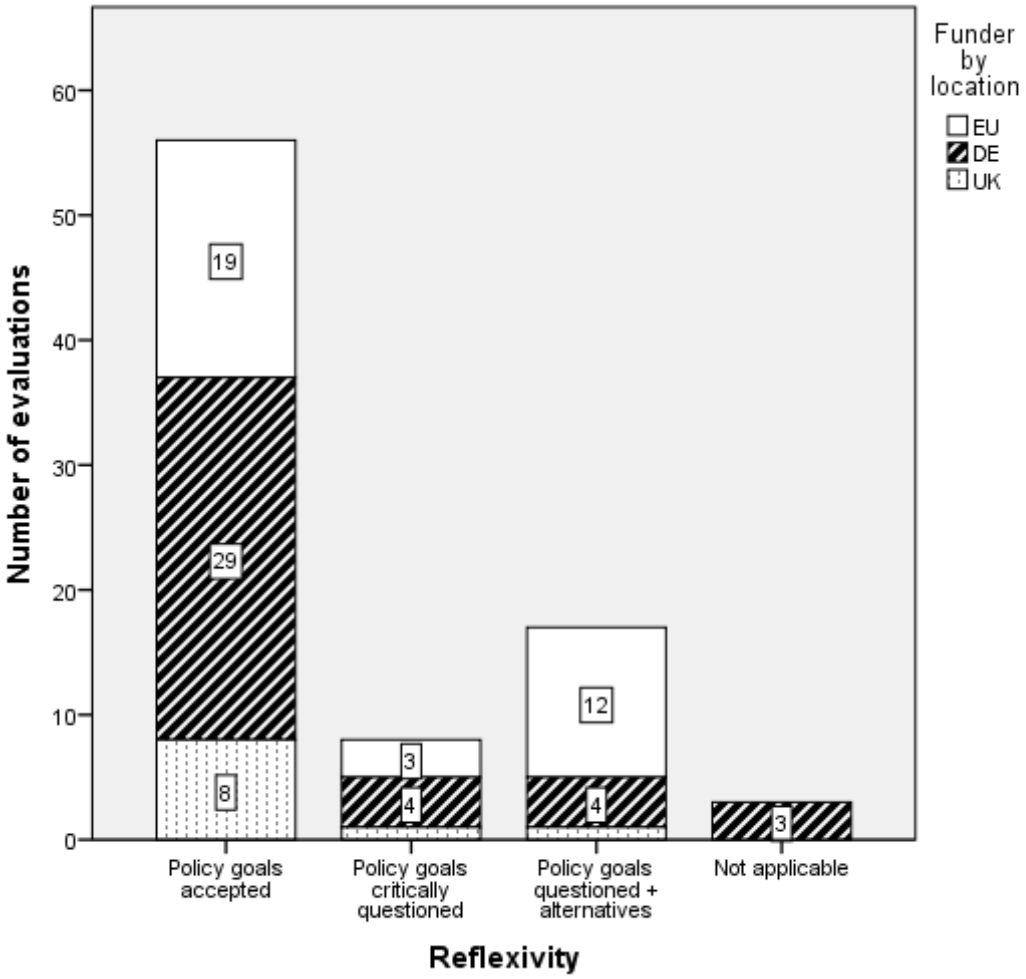
Figure 6.21: Number of criteria in informal evaluations[44]



The final contextual category addressed here relates to reflexivity regarding the policy's goals or targets, or, as explained above, to what extent informal evaluations critically engage with the extant policy targets. Figure 6.22 reveals that even among informally-funded evaluations, the vast majority (64.29%) evaluated policies against their stated aims; that is, a very low level of reflexivity. A much smaller number of evaluations (9.52%) critically questioned policy targets, and only 20.24% of the evaluations actually proposed alternative targets. The highest number of non-reflexive evaluations were funded by actors in Germany, followed by actors at the EU level and then the UK. Funders at EU level fund the greatest number of the 'most reflexive' evaluations, followed by funders from Germany and finally from the UK.

[44] Multiple responses possible.

Figure 6.22: Reflexivity in informal evaluations



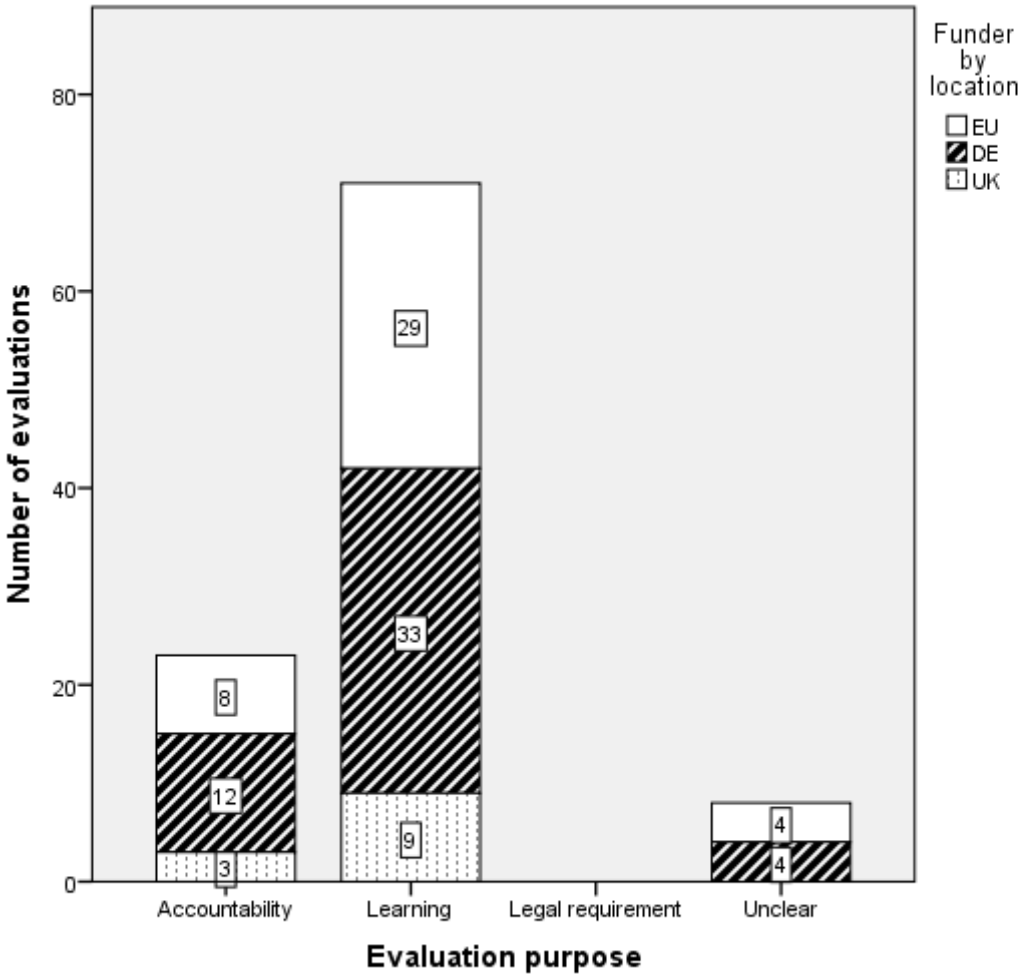
- As Chapter 2 discussed, the empirical evidence confirms that no single dimension addresses context in an unequivocal way. Considerable variation across multiple variables points to the multi-dimensionality of context in evaluation. It is thus not possible to average across all the different dimensions, but some general trends emerge.
- Overall, there appears to be a medium level of attention to context among informal evaluations. Informal evaluation funders from the UK appear to have funded fewer evaluations with higher levels of attention to context than evaluation funders from the EU level or Germany, although there is variation across individual variables.

- The variability contained in these data also points to wide-ranging practices of evaluation, including to the fact that there are, indeed, a range of evaluations that score quite high in some of the relevant contextual dimensions.
- One surprising finding is that informal evaluation are, on the whole, not very reflexive. However, informal evaluators used a wide range of criteria to evaluate climate policies. The following section addresses variables that describe the third foundational idea of polycentrism, namely interaction.

6.4 Interaction

The data presented in this section address the third foundational principle of polycentrism, namely interacting governance centres (see Chapter 2). The first variable addressing this question is the overall evaluation purpose. As Figure 6.23 details, the content analysis reveals a range of reasons for conducting informal evaluations, with almost all evaluations (84.52%) citing learning as one of the core purposes of evaluation. Accountability was cited in only 27.38% of the evaluations analysed here. Fully 9.52% of the informal evaluations did not explicitly state their purpose—this could be an indication of other motivations to evaluate, including the political ones (see Chapter 2). As discussed in the previous section, no informal evaluation responded to a legal requirement to evaluate. The distribution by evaluation funder (shading of the bars in Figure 6.23) is about proportionate.

Figure 6.23: Evaluation purpose[45]

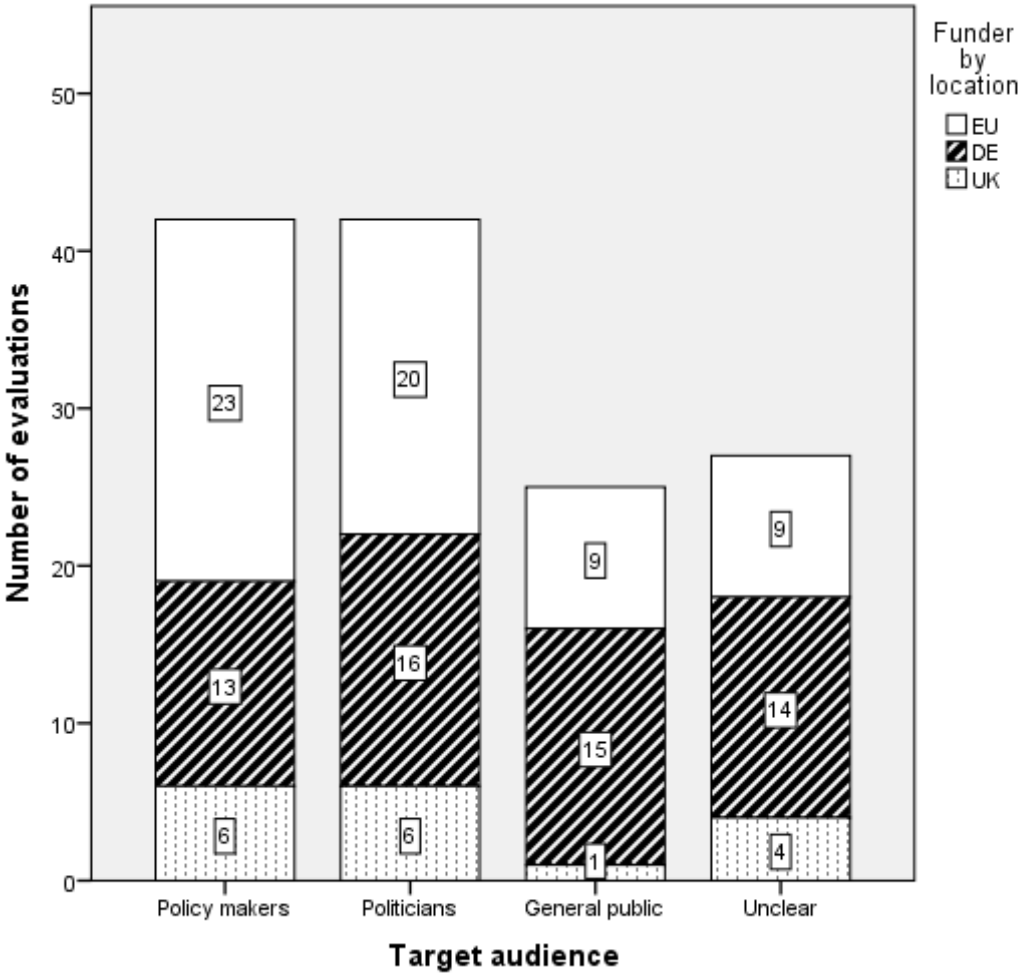


Another view of addressing interacting governance centres is the stated or implied target audience of the evaluation, that is, for whom is the evaluation written? Figure 6.24 reveals that about half of the informal evaluations state that they are geared towards politicians and/or policy makers (recall that the latter are ministry officials or similar). A smaller number of evaluations is geared towards the general public, but it is also notable that for a sizeable number of evaluations, it was not possible to determine the target audience, because these evaluations contain no explicit statement on the target audience. UK based funders financially supported few evaluations for the general public, while Germany based evaluators supported a

[45] Multiple mentions possible.

bigger number (15 out of 84 evaluations). At the same time, it was not possible to determine the target audience for 14 of the evaluations supported by German funders.

Figure 6.24: Target audience[46]

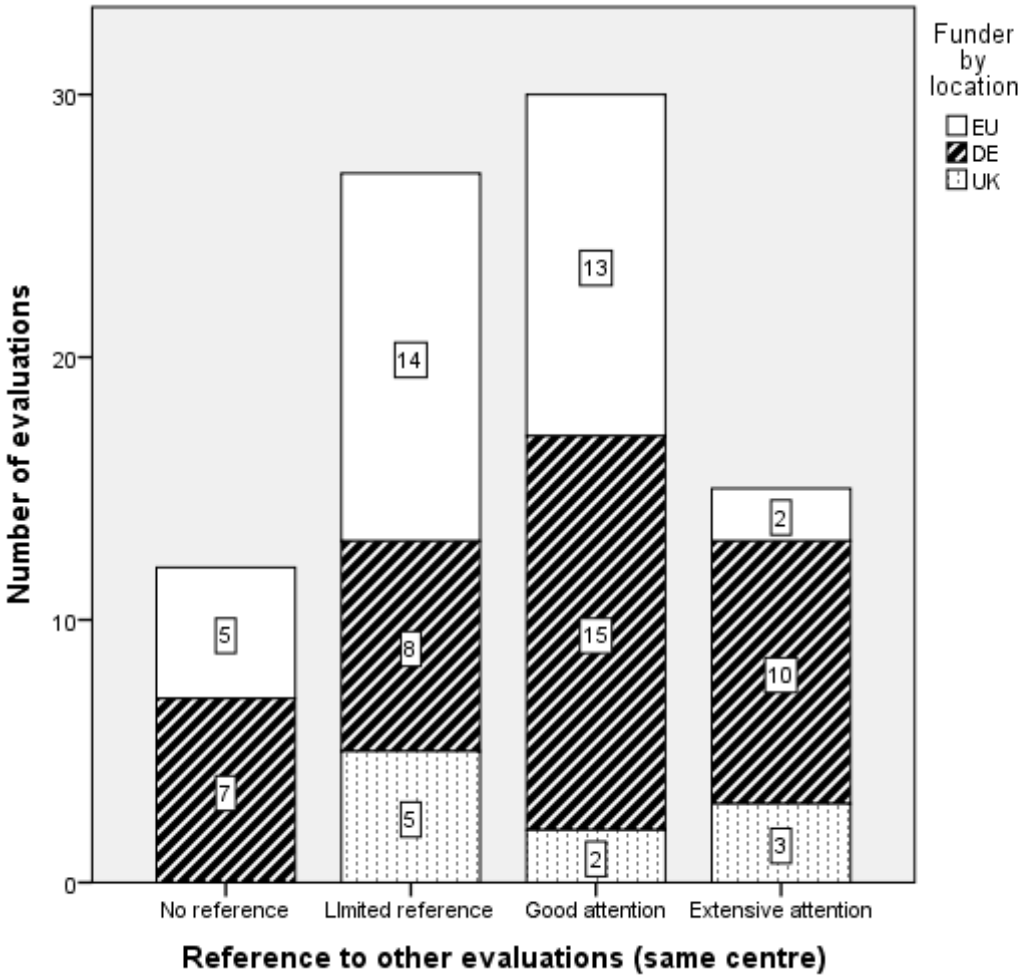


In order to probe interactions with other governance centres more directly, Figure 6.25 presents data on the extent to which the evaluations referred to other evaluations focusing on the same centre (if for example a study focusing on renewables policy in Germany also cites or engages with other studies focusing on renewables or other policies in Germany). Figure 6.25 shows that 53.57% of the evaluations pays either good or extensive attention to other studies focusing on the same centre. By the same token, 32.14% of the evaluations pays limited attention to

[46] Multiple mentions possible.

other evaluations of the same centre and 14.29% pays no attention to other evaluations. Looking at the location of the evaluation funders (shading of the bars in Figure 6.25) reveals that all evaluations funded by informal actors in the UK make reference to evaluations of the same centre. Also, the majority (62.50%) of evaluations funded by actors from Germany pays either good or extensive attention to other evaluations of the same centre or policy. By contrast, actors at EU level funded only two evaluations that pay extensive attention to other evaluations of the same centre.

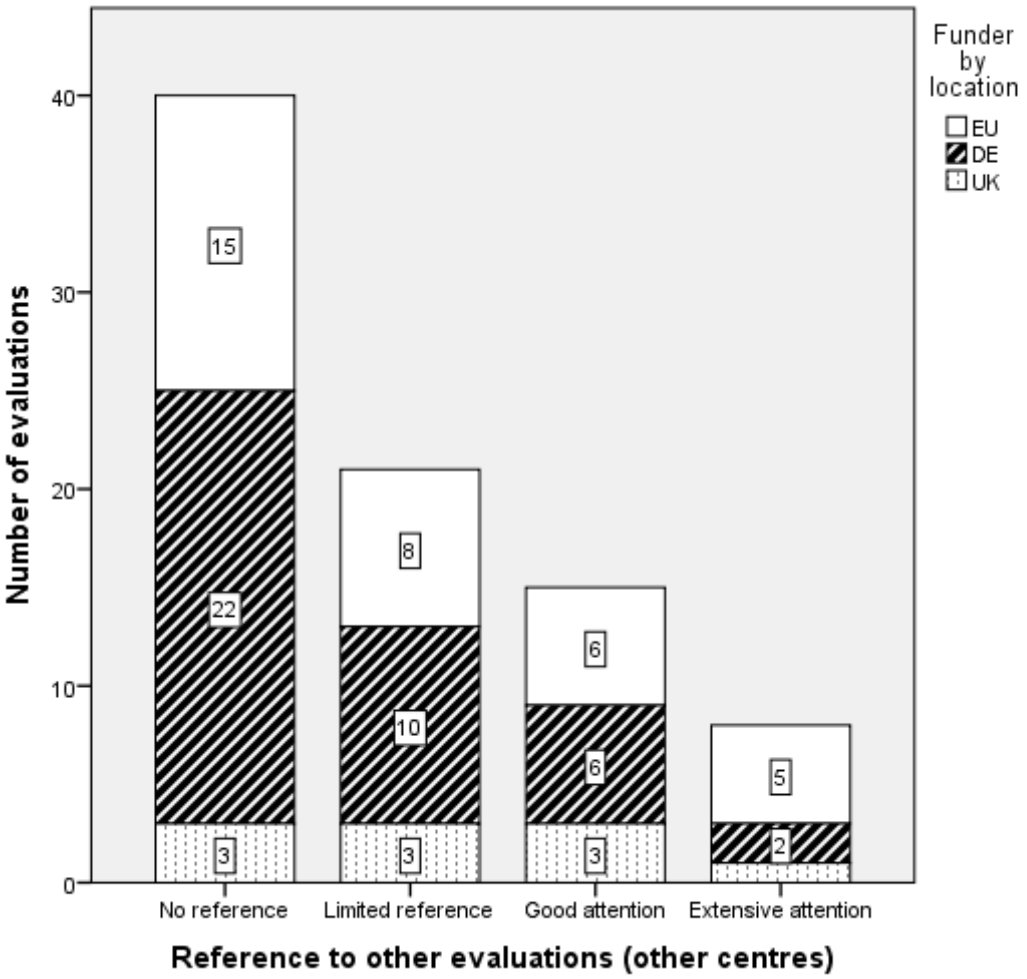
Figure 6.25: References to other evaluations focusing on the same centre



Another way of looking at interacting governance centres in the evaluations is to consider to what extent they discuss experiences from other governance centres.

This would be the case, for example, if an evaluation assessing German renewables policies also took into account experiences with similar policies in France. Figure 6.26 shows that 47.62% of the informal evaluations make no reference to experiences from other centres, and the number of evaluations with either good or extensive attention comprised 27.30% of all informal evaluations. Turning to the distribution by location of the evaluation funders reveals by and large a proportionate distribution across funders.

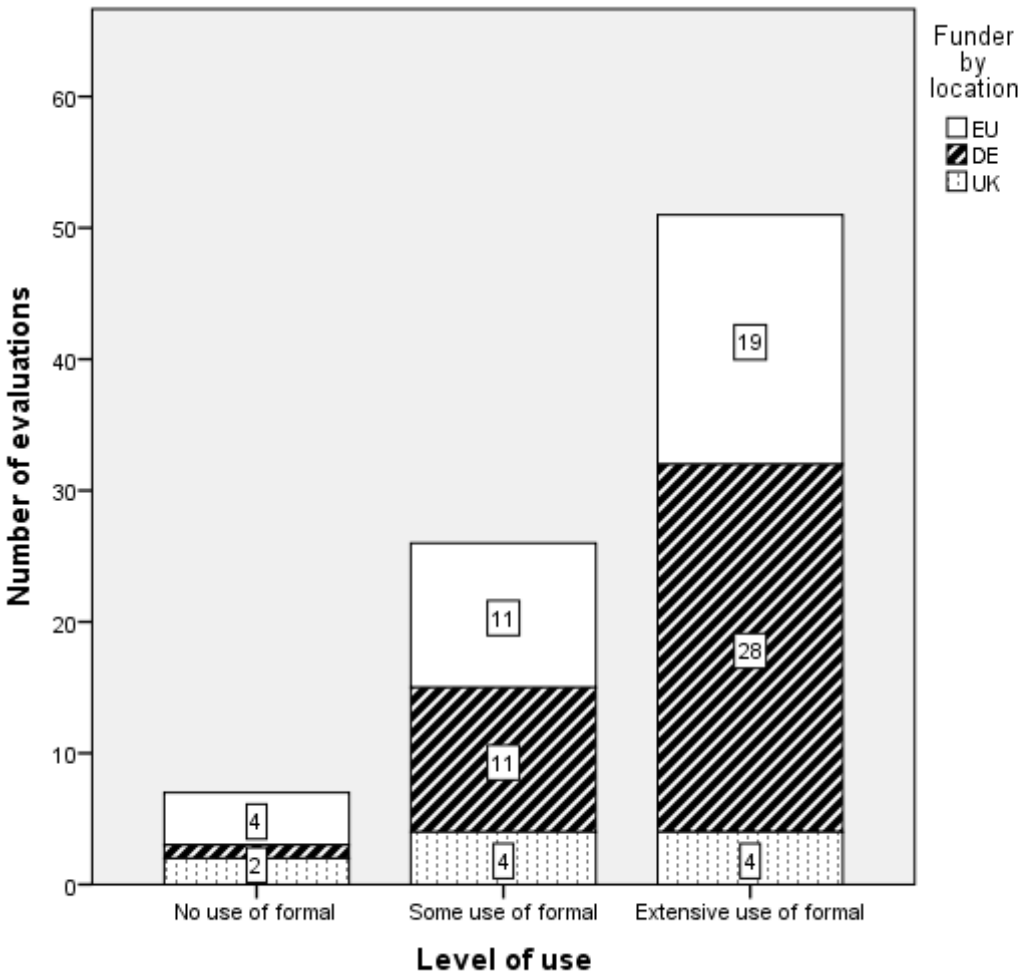
Figure 6.26: References to other evaluations focusing on other centres



Furthermore, Figure 6.27 examines the extent to which informal evaluations draw on insights or data from formal evaluations. The results show that a strong majority of 60.71% of the informal evaluations make extensive use of insights or data from formal evaluations. A further 30.95% of the informal evaluations make

some use of insights or data from formal evaluations, and only 8.33% of the informal evaluations make no use of insights or data from formal evaluations at all. The distribution by evaluation funders (shading of the bars) is by and large proportional (considering the overall distribution in informal evaluation funders discussed above).

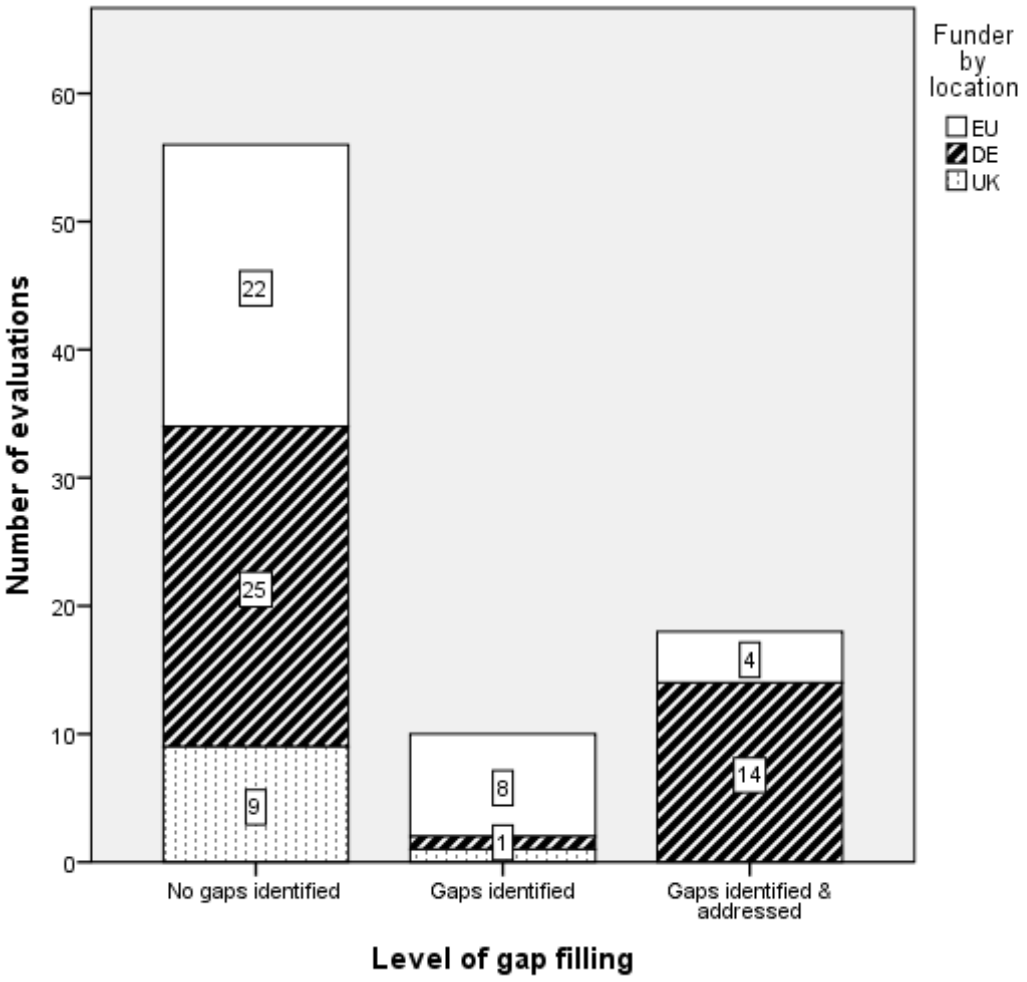
Figure 6.27: Use of insights or data from formal evaluations



In a related vein, Figure 6.28 summarizes data on the extent to which informal evaluations intend to spot and address any gaps left by formal evaluations. An overwhelming majority of informal evaluations (66.67%) make no effort whatsoever to identify gaps in formal evaluation; 11.90% of the informal evaluations spot some gaps in formal evaluation, but only 21.43% of informal evaluation spot *and* address any gaps left by formal evaluations. Considering the distribution by evaluation funder (the shading of the bars) shows that informal evaluations that spot *and* address

gaps left by formal evaluations were mainly funded by actors from Germany (14 evaluations) and to a lesser degree by actors at the EU level (4 evaluations). EU level actors also funded a number of evaluations (8) that identify gaps, but do not address them. Informal actors in the EU funded almost no evaluations that spot, let alone address, gaps in formal evaluations.

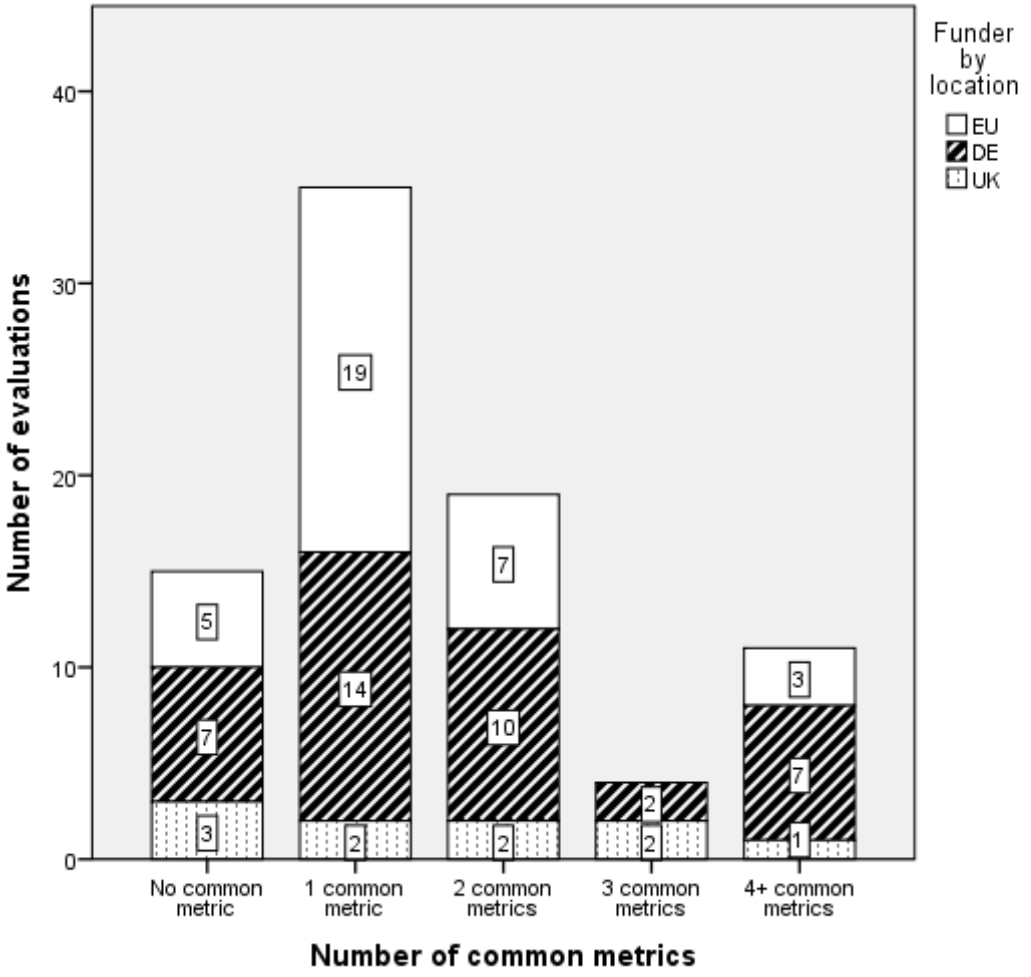
Figure 6.28: Informal evaluation filling gaps left by formal evaluation



One way to link evaluation findings between different governance centres is through comparable quantitative metrics that allow comparison of findings such as greenhouse gas emissions or costs per climate policy. As Figure 6.29 details, most evaluations contain either one, two or more common metrics that could, at least in principle, enable comparison with other studies. The distribution by location of the evaluation funder (shaded bars) is by and large proportionate, although informal actors from Germany exhibit a somewhat greater propensity towards funding

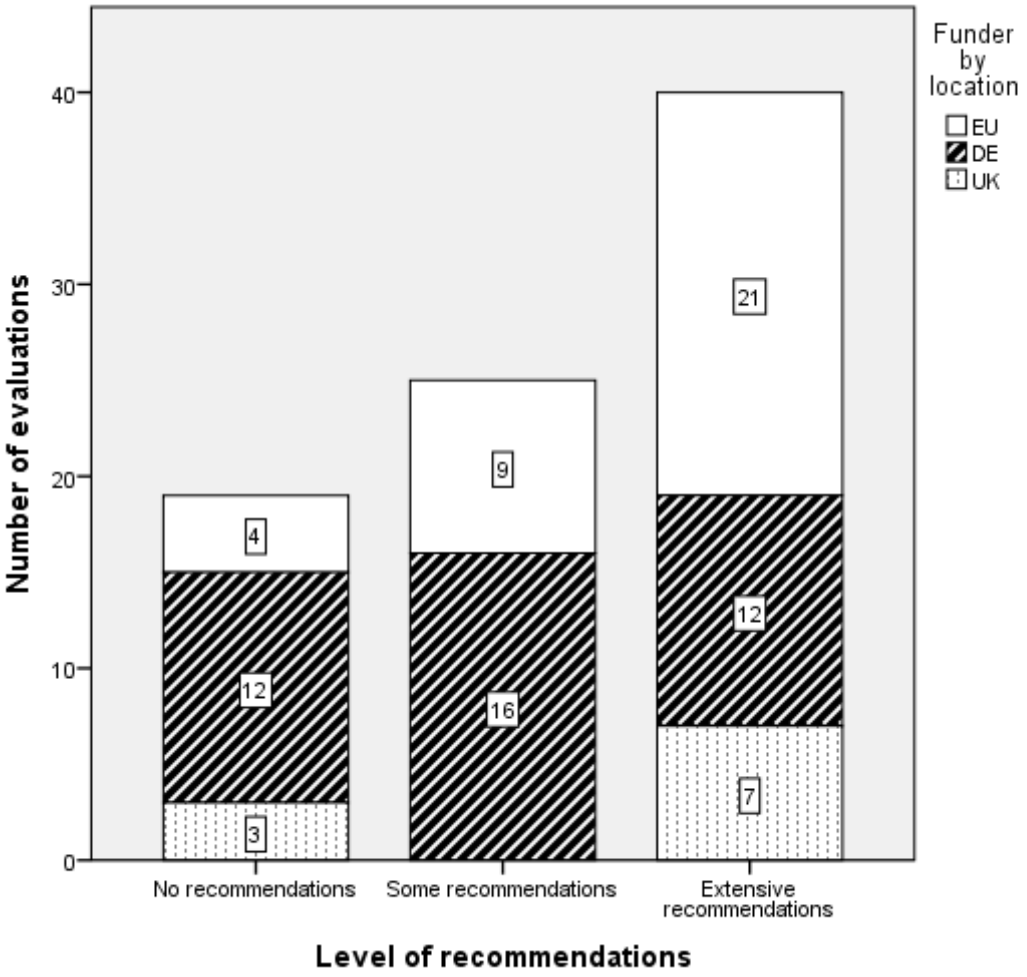
evaluations with a higher number of comparability metrics than funders from the EU level or from the UK.

Figure 6.29: Number of comparability metrics in informal evaluations



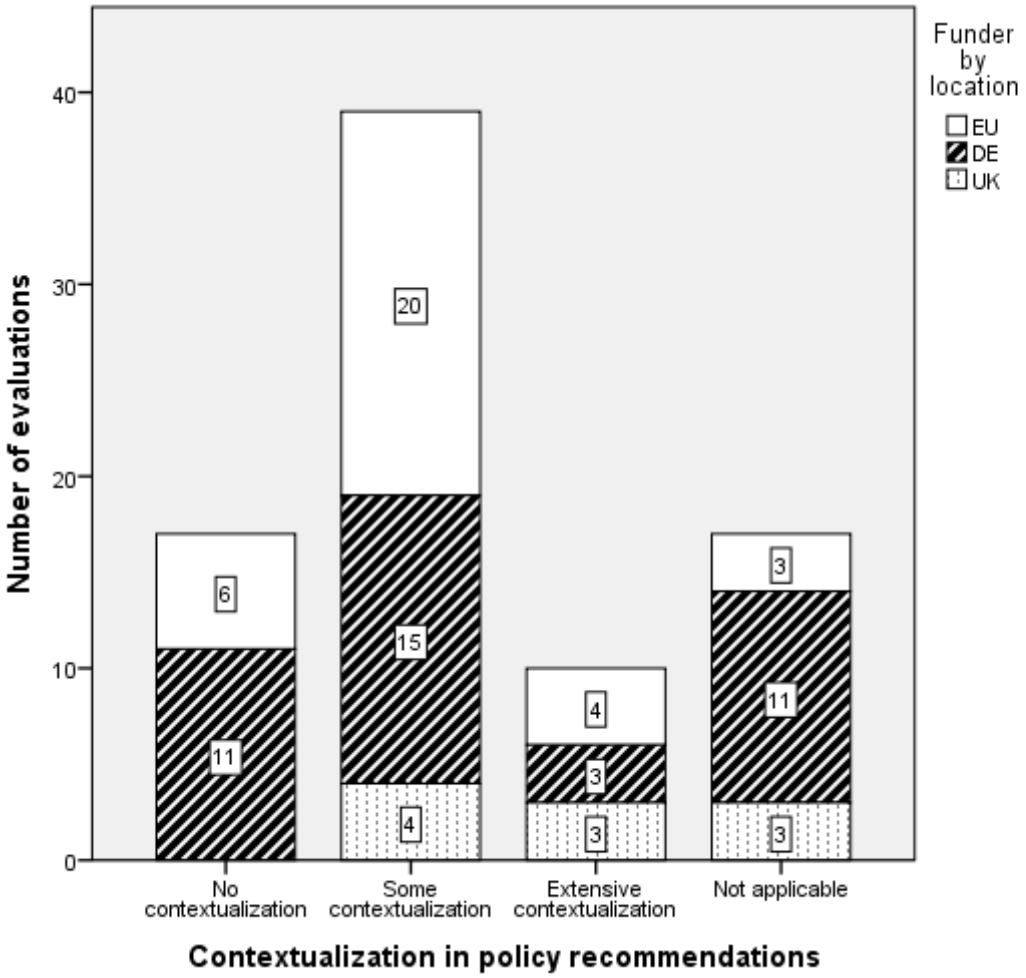
Another way in which other governance centres may directly benefit from an evaluation is through identifying lessons or making recommendations (this also applies to learning effects within the same centre). Figure 6.30 details how most evaluations boast either some or extensive recommendations. Interestingly, informal evaluation funders based in the UK either fund evaluations that provide no recommendations, or extensive recommendations. The biggest number of evaluations funded by actors at the EU level (21 out of 84) exhibit extensive recommendations.

Figure 6.30: Recommendations in informal evaluations



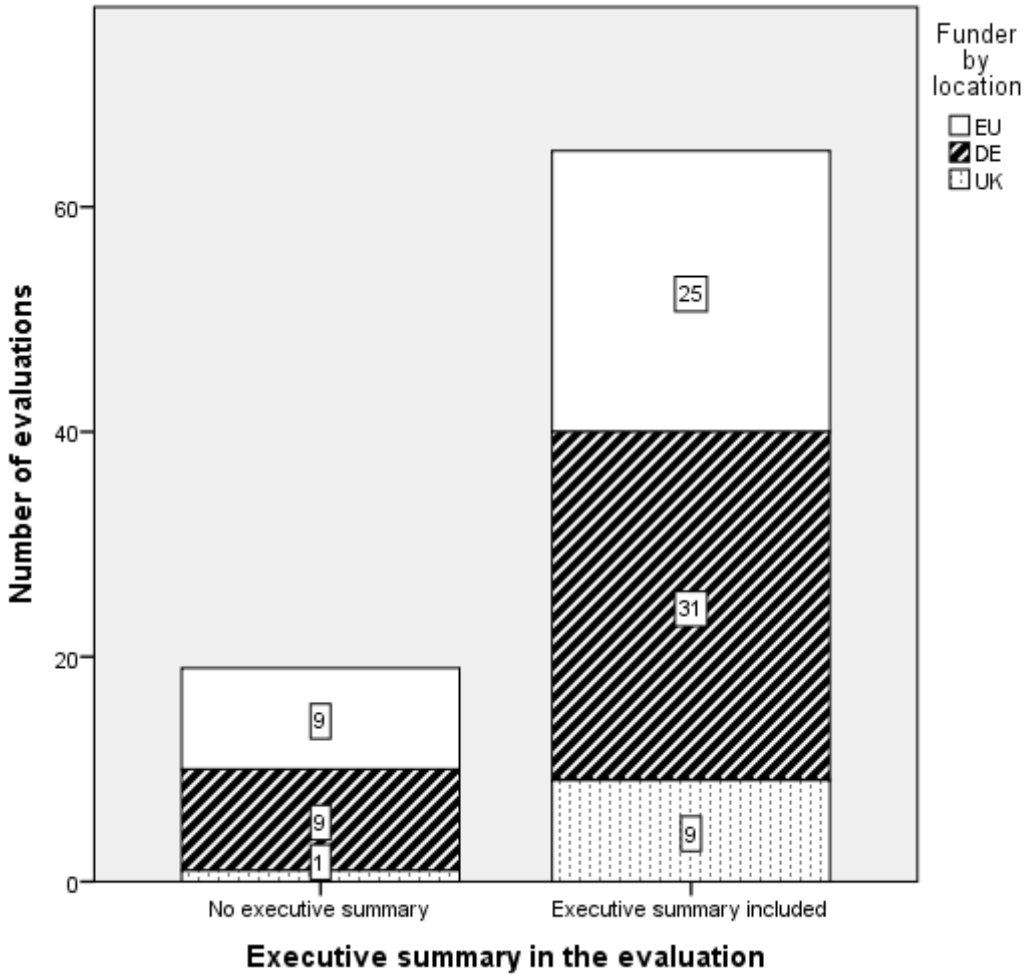
However, Chapter 2 noted that recommendations alone may not be enough to enable drawing lessons from one centre to another; detailed contextual information may indeed be necessary to gauge the transferability of lessons. Therefore, Figure 6.31 details to what extent the informal evaluations make recommendations and contextualize them. The data show that 58.33% of the evaluations exhibit either some or extensive contextualization of the recommendations (which refers to making references to the various contextual dimensions reviewed above), but for the other half this is either not applicable (i.e. no recommendations) or the recommendations are not contextualized at all. Looking at the distribution by evaluation funder (the shading of the bars) reveals that UK based actors only fund evaluations that contextualize their recommendations; in the category 'extensive contextualization', the distribution among evaluation funders is about even.

Figure 6.31: Contextualization of policy recommendations



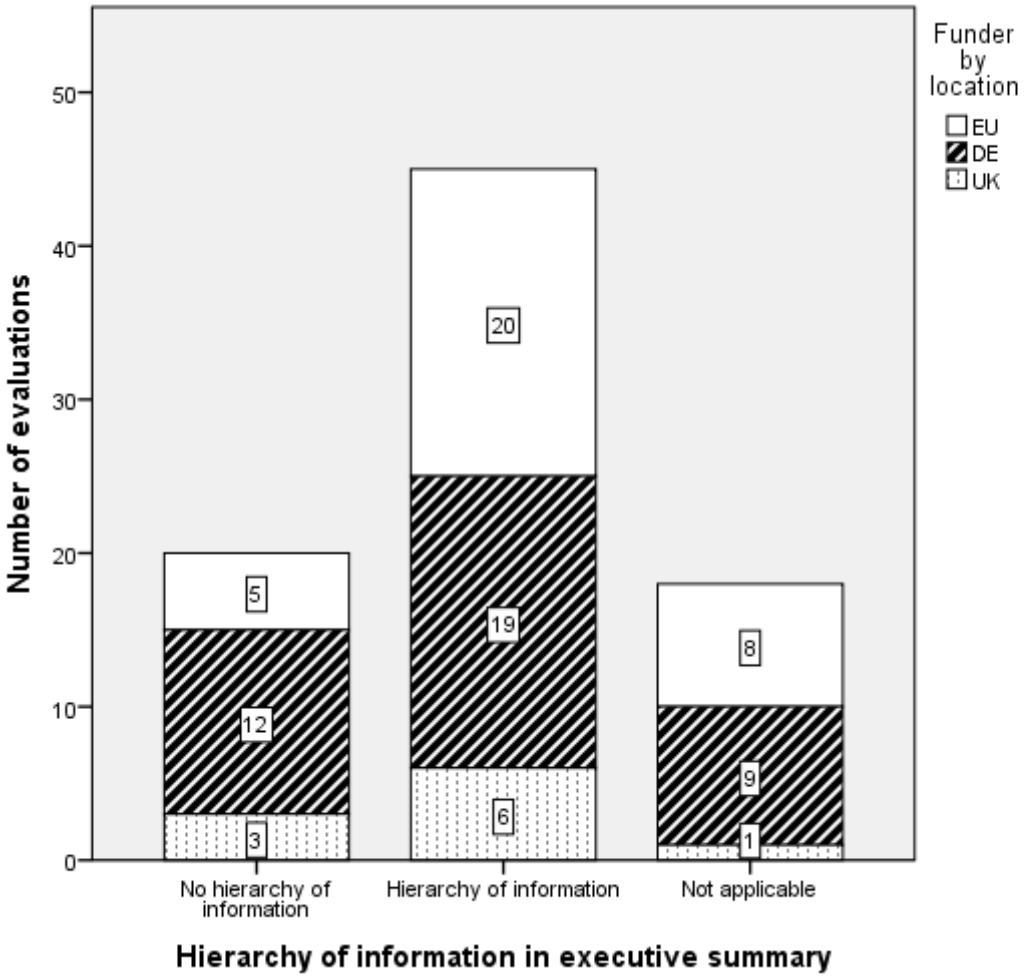
Another important way of gauging the extent to which evaluations may actually be designed to carry lessons is their own structure. An important element is an executive summary, which allows a quick overview of the main findings (see Chapter 2). Figure 6.32 details that 77.38% of the informal evaluations include an executive summary. Looking at the distribution by evaluation funder reveals that UK based informal actors funded only one evaluation that does not include an executive summary, while both actors at the EU level and from Germany funded nine evaluations each that contain no executive summaries.

Figure 6.32: Executive summaries in informal evaluations



An executive summary may still fail to engage busy policy-makers, unless the information contained therein is clearly structured. Figure 6.33 thus describes data on the extent to which the information in the executive summary exhibits some sort of structure, be it through bullet points, graphs, tables, bolding, or other means. A majority (53.57%) of the evaluations contain some sort of hierarchy of information in their executive summary, while for 23.81% the inverse is true. The distribution by evaluation funder is about proportionate, although UK based informal actors exhibit a somewhat lower tendency to fund evaluations that do not include a hierarchy of information in their executive summaries.

Figure 6.33: Hierarchy of information in executive summaries



In order to enable exchange across different governance centres, particularly in the multi-cultural and multi-lingual environment of the EU, accessibility to evaluation findings in different languages is a key concern. Figure 6.34, however, reveals that the vast majority of informal evaluations (80.95%) contains no summary or even a translation into another language, which remains relatively rare and only appears in about a fifth of the informal evaluations. Looking at the evaluations that do contain a translation or summary in another language shows that informal funders based in Germany had funded most evaluations with this characteristic (12), followed to a considerably lesser extend by evaluation funders at the EU level (3 evaluations) and an even lower number from the UK (one evaluation).

Figure 6.34: Linguistic access to evaluation by funder location

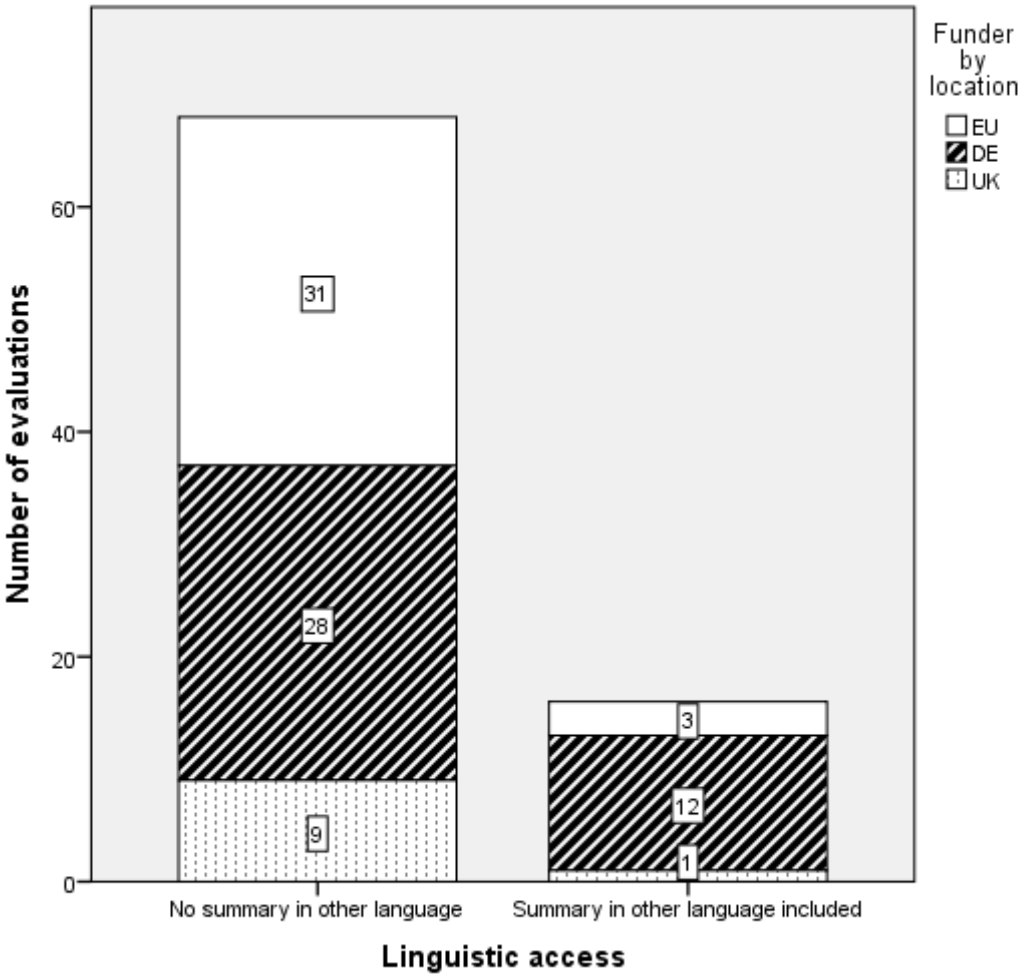
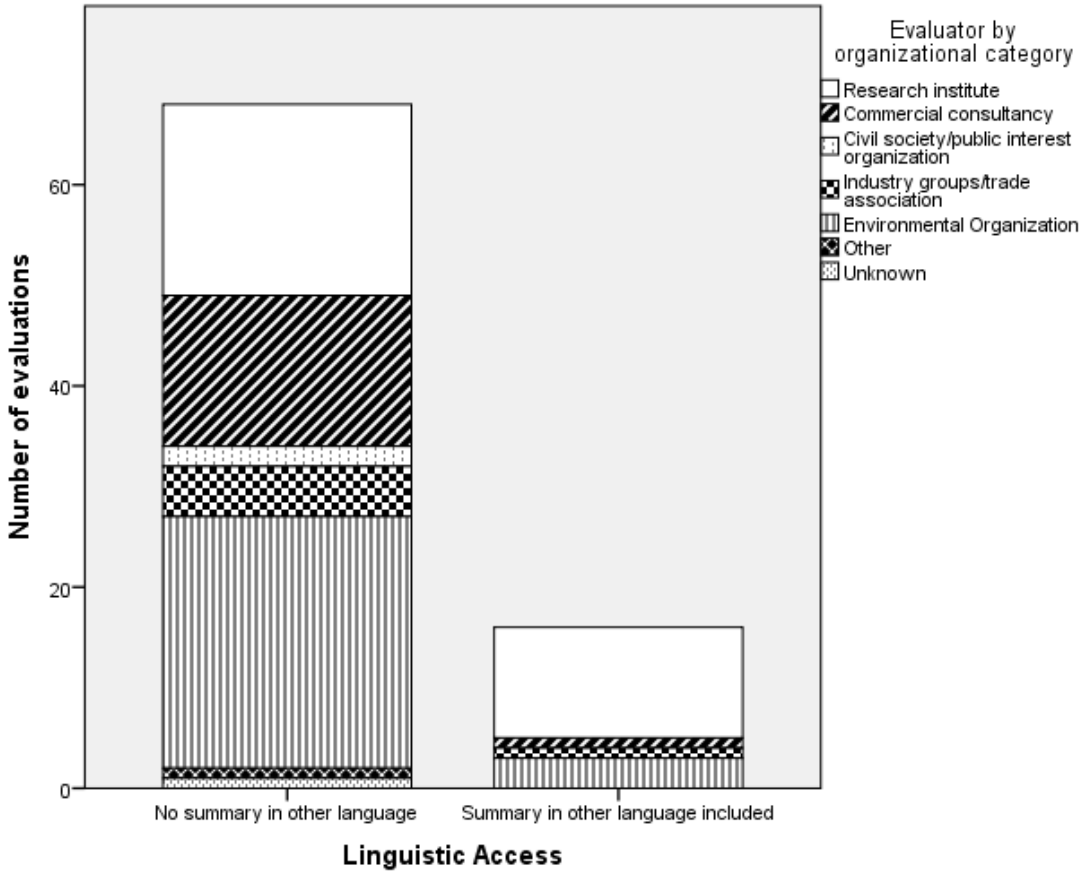


Figure 6.35 presents the same data on linguistic access split up by the organizational category of the organization that conducted the evaluation (the evaluator). Here, we see that evaluations with a summary in another language were almost exclusively conducted by either research institutes or by environmental organizations.

Figure 6.35: Linguistic access to evaluation by evaluator organization type

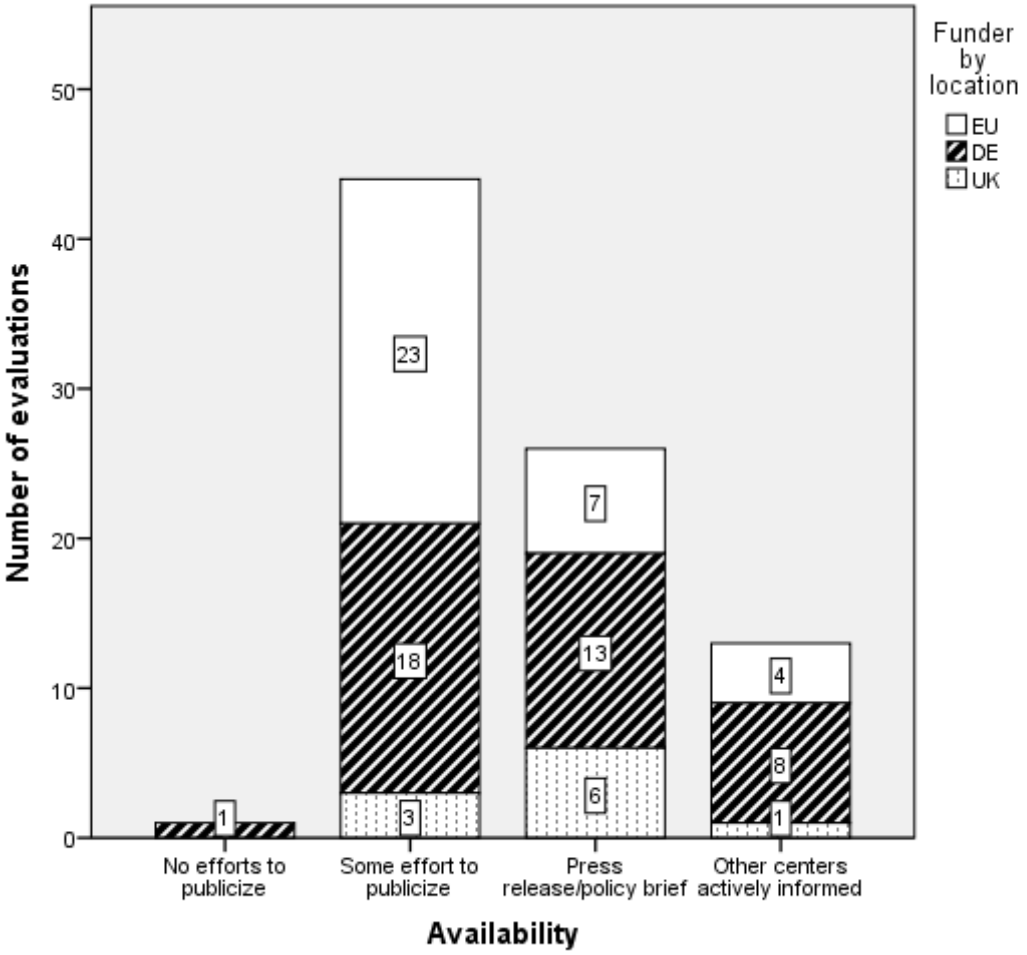


Finally, Figure 6.36 describes data on the availability of evaluation: almost all informally-funded evaluators endeavoured to publicize their evaluations, with more than half including some publicity.[47] However, by the same token, only 15.48% of the informally-funded evaluation authors actively informed other governance centres of their evaluations. The distribution by the location of the evaluation funders is relatively even, although evaluations where other governance centres were actively informed were mainly funded by informal actors in Germany (8 evaluations),

[47] Recall that this was coded on the following scale: 0 = published without any efforts to publicize; 1 = some effort to publicize (e.g., available on website); 2 = press release/policy brief; 3 = other governance centres actively informed about the evaluation outcomes (e.g., press conference, available in multiple outlets, picked up by the media).

followed by actors at the EU level (4 evaluations) and, finally, from the UK (only one evaluation).

Figure 6.36: Evaluation availability



- With a view to understanding the interaction between governance centres, informal evaluations from all three governance centres by and large covered some of the more common ways, such as providing executive summaries or making some references to insights from other evaluations of the same governance centres well.
- However, turning to the deeper ways of more wide-ranging interactions, such as linking with insights from other governance centres, providing translations of findings, including quantitative comparability metrics or actively informing governance centres, the informal evaluations did less well.

- From a polycentric perspective, this indicates that informal, self-organizing actors contribute some important ways in which governance centres may interact, but do not cover all areas equally. The ability of informal evaluations to cover gaps left by formal evaluation is thus limited.

6.5 Conclusion

The analyses presented in this chapter reveal that it is possible to disaggregate the 'informal' evaluation category in order to shed more light on the activities of non-state actors. The first thing to note is that in conjunction with the findings from the overall database presented in Chapter 4, self-organizing capacities in climate policy evaluations are limited in numbers. But where self-organization does happen, environmental groups take a leading role in both funding and ultimately conducting climate policy evaluations, while research institutes and private consultancies also contribute as evaluators. The finding that informal actors do not fund any formal ones to conduct evaluations is desirable from a democratic perspective - it would be highly problematic if for example a private company paid a court to conduct an evaluation.

Turning to the content analysis of the evaluations, it is clear that self-governance, context, and interacting governance centres are multi-faceted concepts that cannot be measured with one or just very few variables. While some variables may be combined in indexes (such as the context index calculated above), others remain more stand-alone aspects. Overall, the evidence presented here suggests that informal evaluations engage with contextual factors to a moderate degree, and they provide moderate ways to stimulate interactions between governance centres. From a polycentric governance perspective (see Chapter 2), it is somewhat surprising that only few informal evaluations engage crucially with extant policy goals (i.e. exhibiting reflectivity), given that informal actors are often thought to be more independent and thus potentially more reflexive.

Finally, this chapter presents each set of variables and analysis in separate sections. The following Chapter 7 will compare the findings from the formally and informally funded evaluations (i.e. the findings that Chapters 5 and 6 presented) in

greater depth, while Chapter 8 discusses potential overlaps between the different variables with a view to the three foundational ideas.

Chapter 7 Comparing Formal and Informal Evaluation

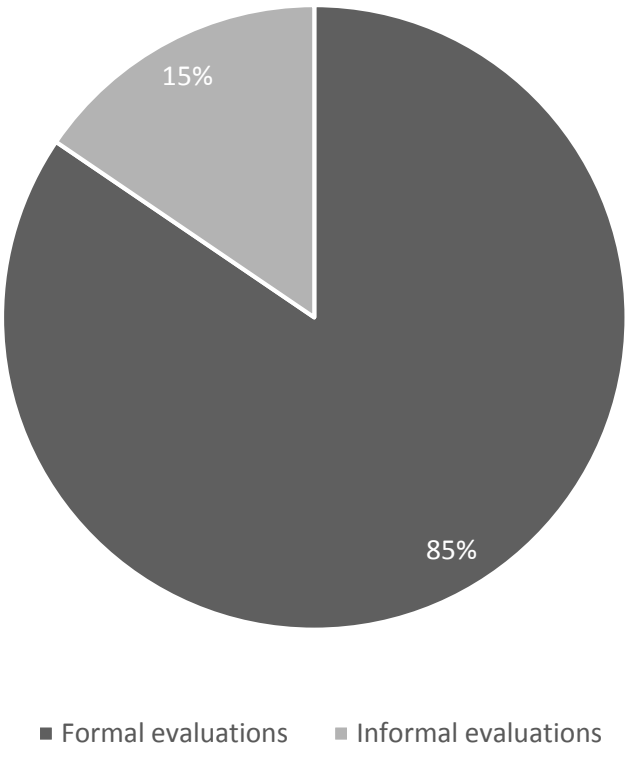
7.1 Introduction

This chapter compares formal and informal climate policy evaluations. It therefore compares the empirical findings presented in Chapter 5 and Chapter 6 in order to work towards a more systematic empirical understanding of climate policy evaluation. Recall that by definition (see Chapter 4), formal evaluation funders include courts and scrutiny bodies, parliaments, governmental organizations, banks, or agencies, independent advice committees, research councils, research institutes/ universities, and government (policy-makers). By the same token, informal evaluation funders include environmental organizations, foundations, industry, and public interest organizations. Given that Chapters 5 and 6 have already explored data on formal and informal evaluations individually (and in that context, looked at the evaluations funded by actors at EU level, in Germany and in the UK), the following sections compare formal and informal evaluations along the foundational ideas of polycentric governance, namely self-organization, context, and interaction (see Chapter 2). Some of the figures below significantly expand on those presented in Chapter 4 (Section 4.4.4). But while the earlier figures focus on the governance centre of the evaluation funders (i.e. EU level, Germany and the UK), the current section focuses on the difference between formal and informal evaluations, given the high relevance of these dimensions in the current thesis. The comparisons draw on a mixture of descriptive, graphic analysis, as well as some statistical comparison where possible and necessary. This chapter concludes with some broader insights emerging from the data with a view to the more detailed theoretical discussion of the findings in Chapter 8.

7.2 Self-organization

In the overall evaluation database (N = 618), there are 458 formally-funded evaluations and 84 informally-funded climate policy evaluations.[48] This means that there are more than five times more formal evaluations than there are informal evaluations. Formal actors have thus funded an overwhelming majority of the *ex-post* climate policy evaluations emerging from the EU level, from Germany, and from the UK. This points to limited self-organizing capacities in the realm of climate policy evaluation in the EU. Figure 7.1 highlights that this translates into 85% formal evaluations and 15% informal evaluations (evaluations whose funder could not be determined are excluded here).

Figure 7.1: Formal and informal climate policy evaluations (N = 542)[49]



[48] For the remainder of the evaluations, the funders could not be determined—see Chapter 4.

[49] Evaluations whose funder could not be determined are excluded here.

But as Chapters 5 and 6 already noted, such aggregate numbers mask a great deal of underlying variability. Therefore, Figure 7.2 plots the number of formal and informal evaluations over time. The evaluations included in Figure 7.2 are all informal evaluations, as well as the random sample from the formal evaluations described in Chapter 5 (total N = 168). The first thing to note is that the numbers of formal and informal evaluations mirror each other, with formal evaluations constantly at a higher level. There is also strong growth in informal evaluations after 2010. Furthermore, the number of formal evaluations has been somewhat steadier over time than the number of informal evaluations, whose numbers have proved more volatile (with the exception of the notable dip in 2009).

Figure 7.2: Formal and informal evaluations over time (N = 168)

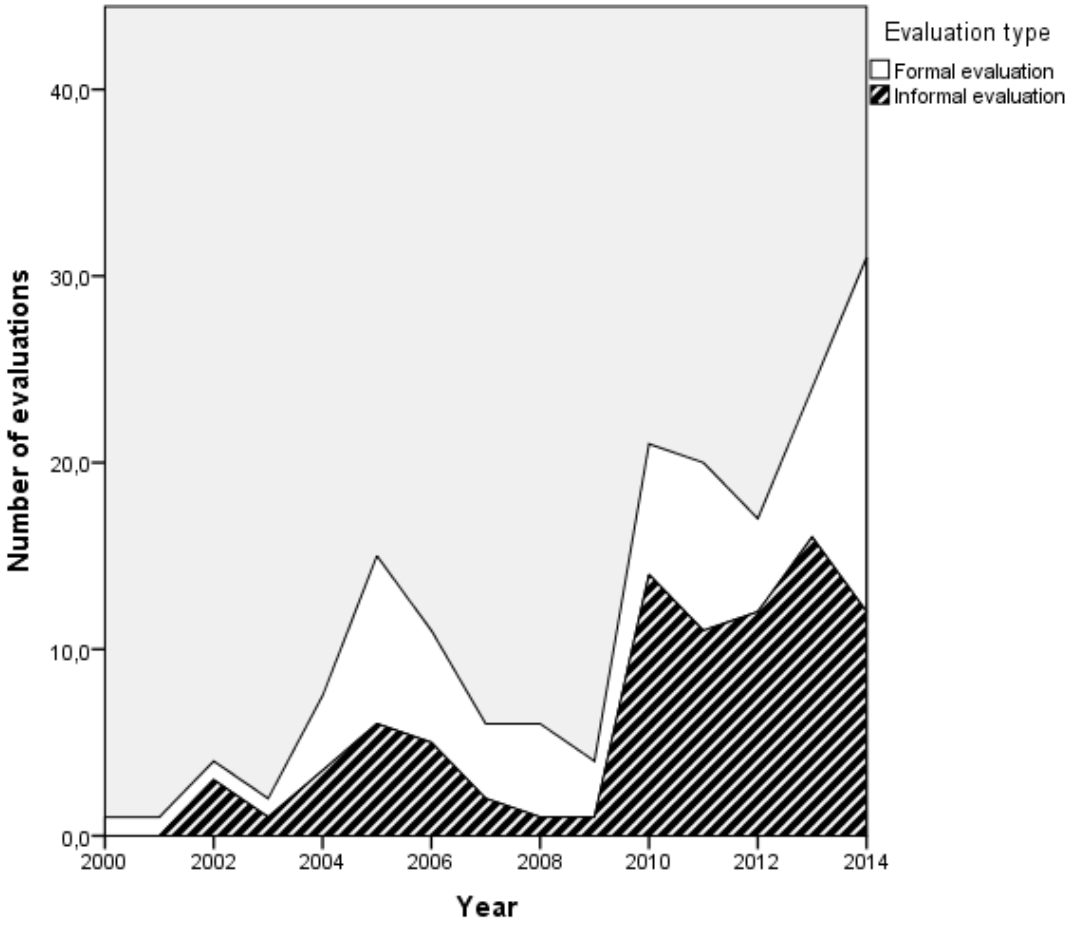
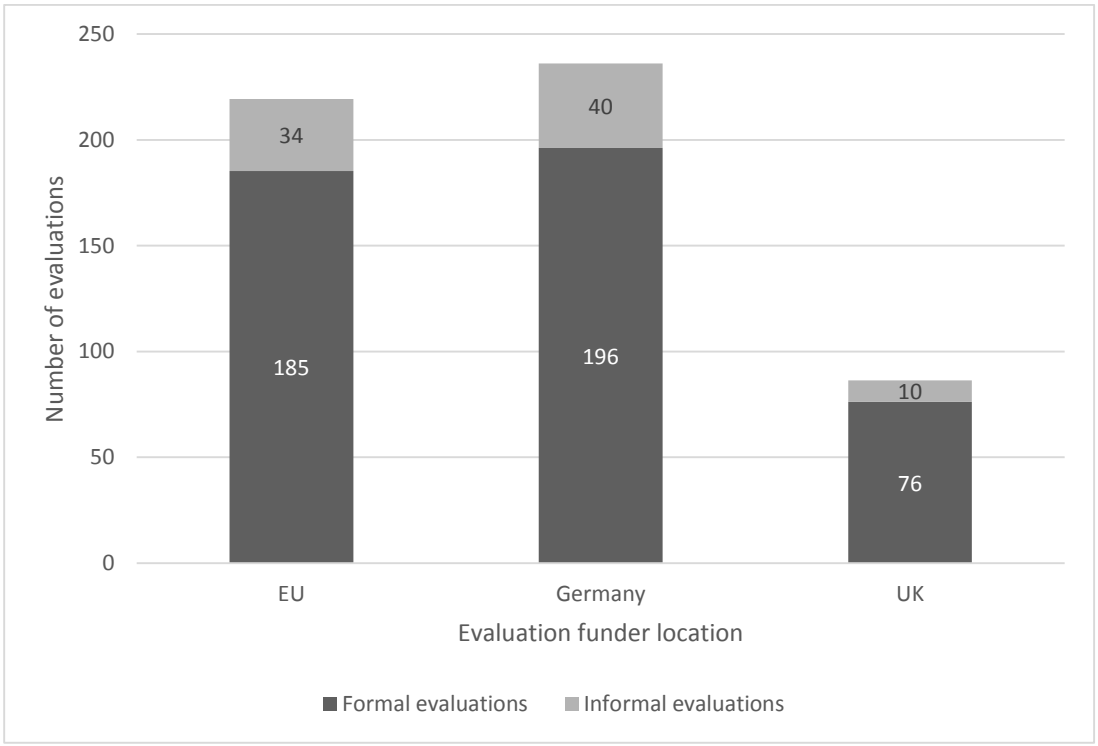


Figure 7.3 represents the number of formal and informal evaluations by the location of the funder (i.e. the EU level, Germany, or the UK). In order to compare

the estimated total of evaluations funded in each governance centre (but recognizing that this thesis analysed a random sample from the formal evaluations—see Chapter 4), the findings that follow incorporate a simple extrapolation by multiplying the sample values for formally-funded evaluations by 5.45 (or the total number of formally-funded evaluations, 458, divided by 84, which is the size of the sample). The same procedure has been applied to a range of data that the figures below present – footnotes clearly denote when this was done. Figure 7.3 reveals that the UK has – based on this estimation and proportionately – the highest ratio of formal to informal evaluations (formal/ informal = 7.6), followed by the EU level (5.4) and then Germany (4.9).

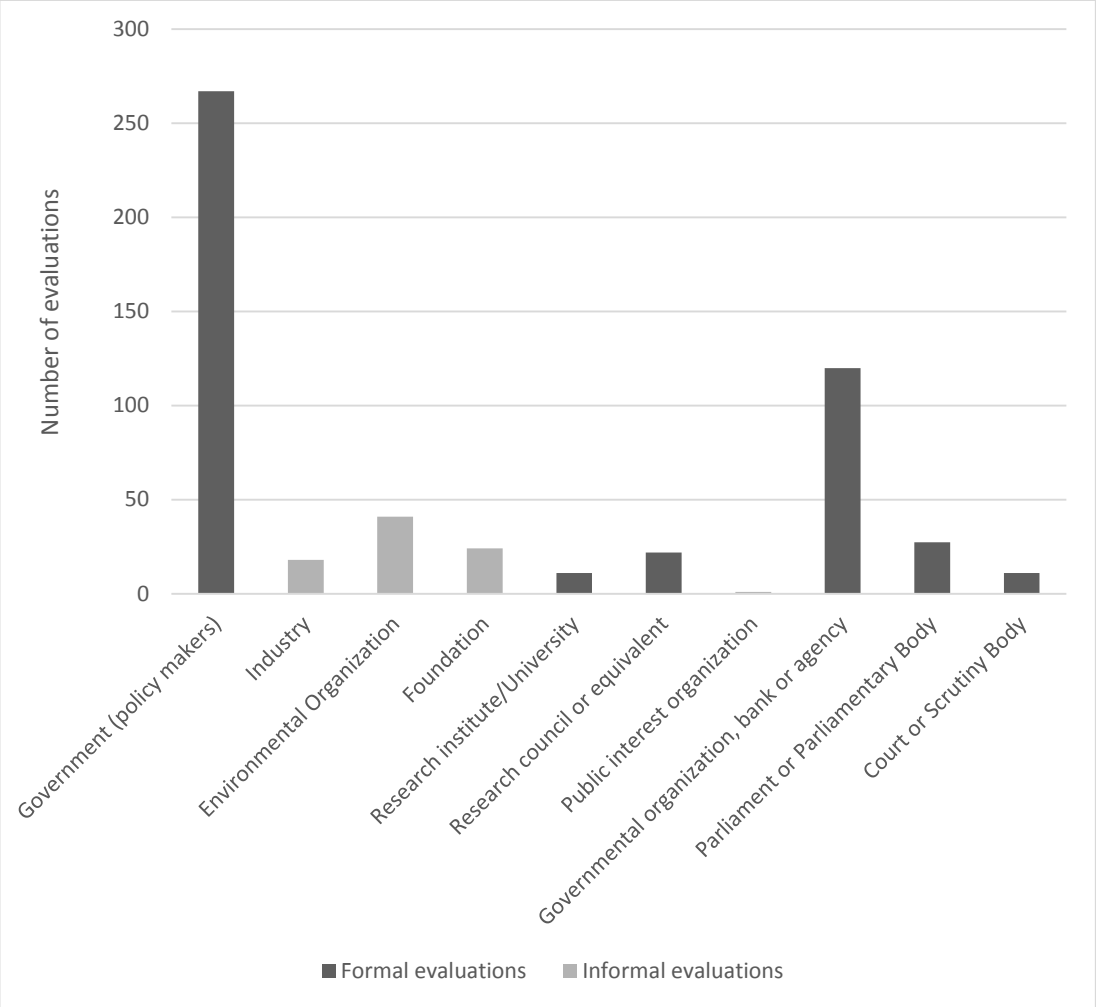
Figure 7.3: Formal and informal evaluations by funder country (N = 542) [50]



[50] Extrapolation for formally-funded evaluations; evaluations whose funders could not be determined are excluded.

Figure 7.4 summarizes these data with a view to understanding the number of evaluations across the organizational type of the evaluation funders. It highlights the strong role of governments and governmental agencies in funding evaluations, while environmental groups funded the greatest number of evaluations in the informal group.

Figure 7.4: Formal and informal evaluations by funder organizational category (N = 542) [51]



[51] Extrapolation for formally-funded evaluations; evaluations whose funders could not be determined are excluded.

Figure 7.5 demonstrates how the formal-informal comparison plays out across by the evaluator location. Given the high level of congruence between the evaluation funder and the location of the evaluator (see Chapters 5 and 6), it is not surprising that the proportions in Figure 7.5 are relatively similar to those in Figure 7.3 above with the exception of just a small number of evaluations conducted by evaluators beyond the EU level, Germany, and the UK.

Figure 7.5: Formal and informal evaluation: evaluator location (N = 542) [52]

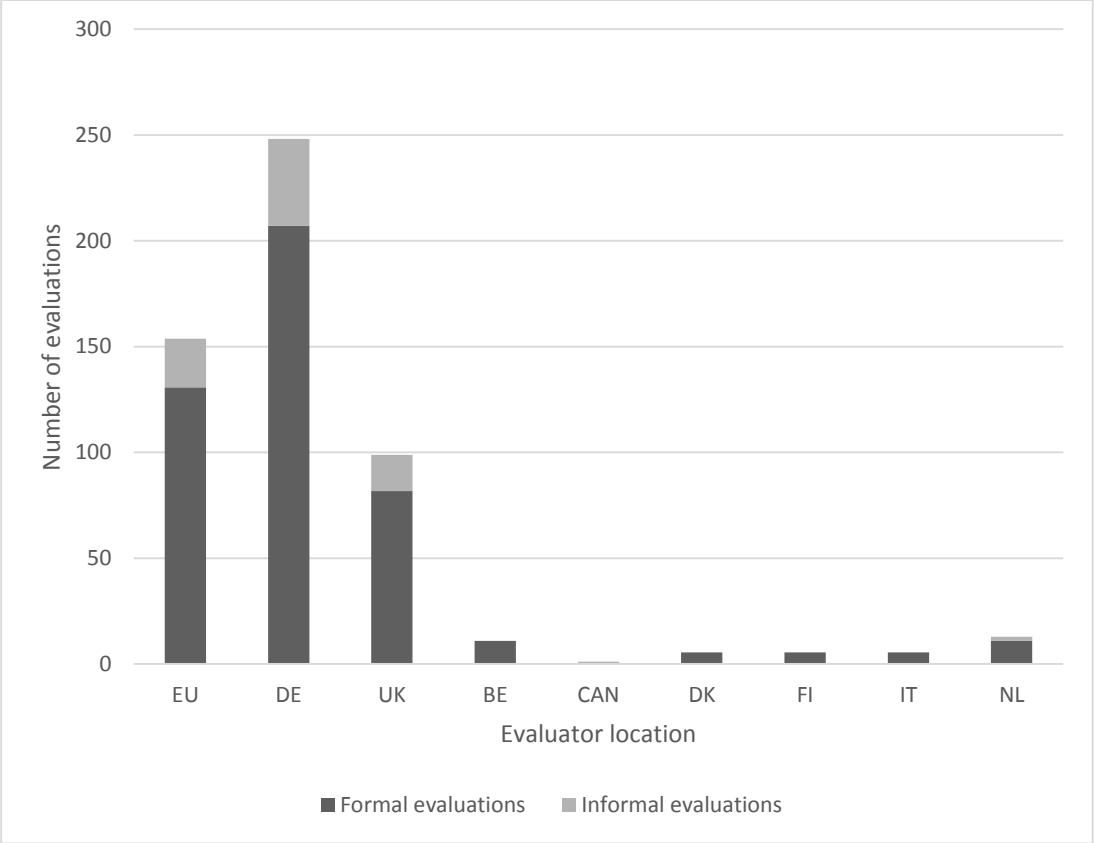
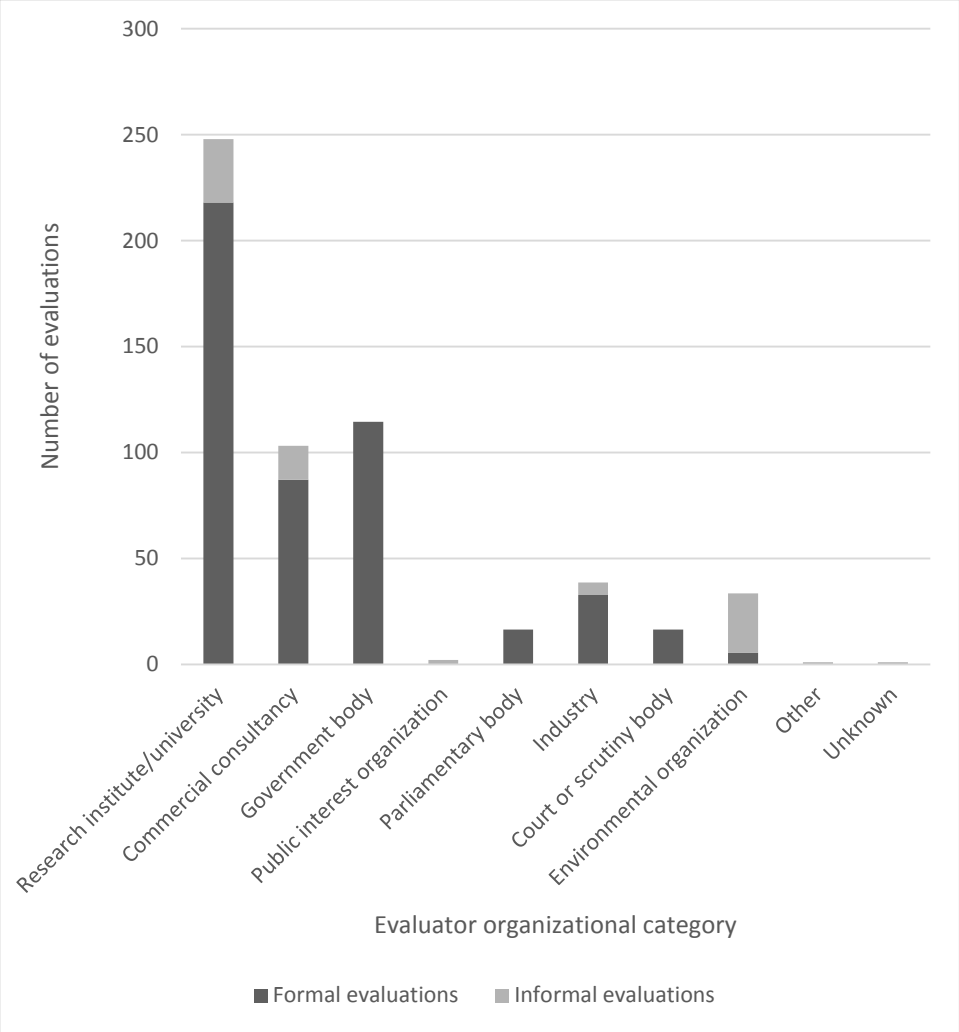


Figure 7.6 then presents the formal-informal comparison with a view to the evaluator types, that is, the organizations that ultimately conducted the evaluations (recall that self-funding is possible). The figure highlights that informal funders

[52] Extrapolation for formal evaluations; evaluations whose funders could not be determined are excluded.

relied by and large on environmental groups, followed by research institutes and commercial consultancies. Formal funders mainly relied on research institutes, governments, and commercial consultancies.

Figure 7.6: Formal and informal evaluations by evaluator organizational type (N = 542) [53]

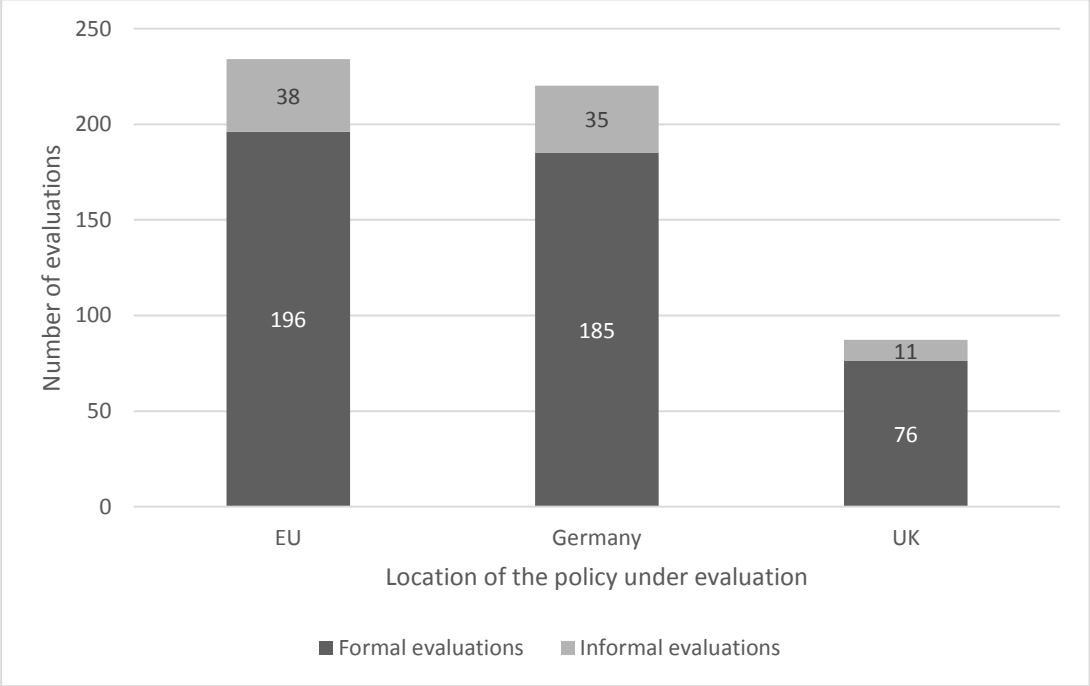


A very similar rationale applies to the location of the climate policy under evaluation in relation to the formal-informal comparison, which Figure 7.7

[53] Extrapolation for formal evaluations; evaluations whose funders could not be determined are excluded.

demonstrates. Again, because evaluation funders tend to fund evaluations of climate policy in their own governance centres (see Chapters 5 and 6), the overall proportions in Figure 7.7 are similar to those in Figure 7.3.

Figure 7.7: Formal and informal evaluation by location of the climate policy (N = 542) [54]

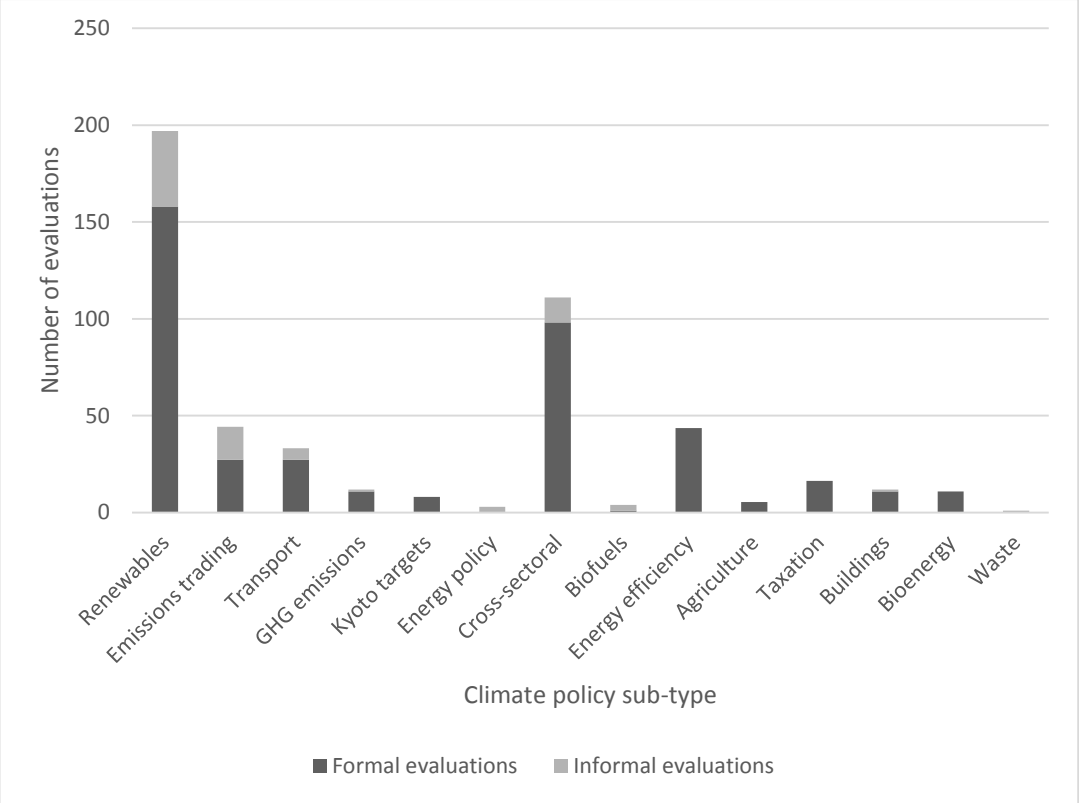


As Chapters 5 and 6 already highlighted, formal evaluation covers a greater number of climate sub-policies than informal evaluation. Figure 7.8 summarizes the respective comparative data. Informal evaluations make a particularly strong contribution when it comes to renewables policy, to emissions trading, as well as to cross-sectoral climate policy. Formal evaluations focus on renewables, cross-sectoral policies, and energy efficiency. One readily visible feature is that, in general terms, the core foci of formal and informal evaluations (i.e. the climate policy sub-sectors with the highest numbers of evaluation) are relatively similar, but there are some

[54] Extrapolation for formal evaluations; evaluations whose funders could not be determined are excluded.

policy areas where there are only formal evaluations (such as energy efficiency, taxation and bioenergy).

Figure 7.8: Formal and informal evaluations by climate policy sub-type (N = 542) [55]

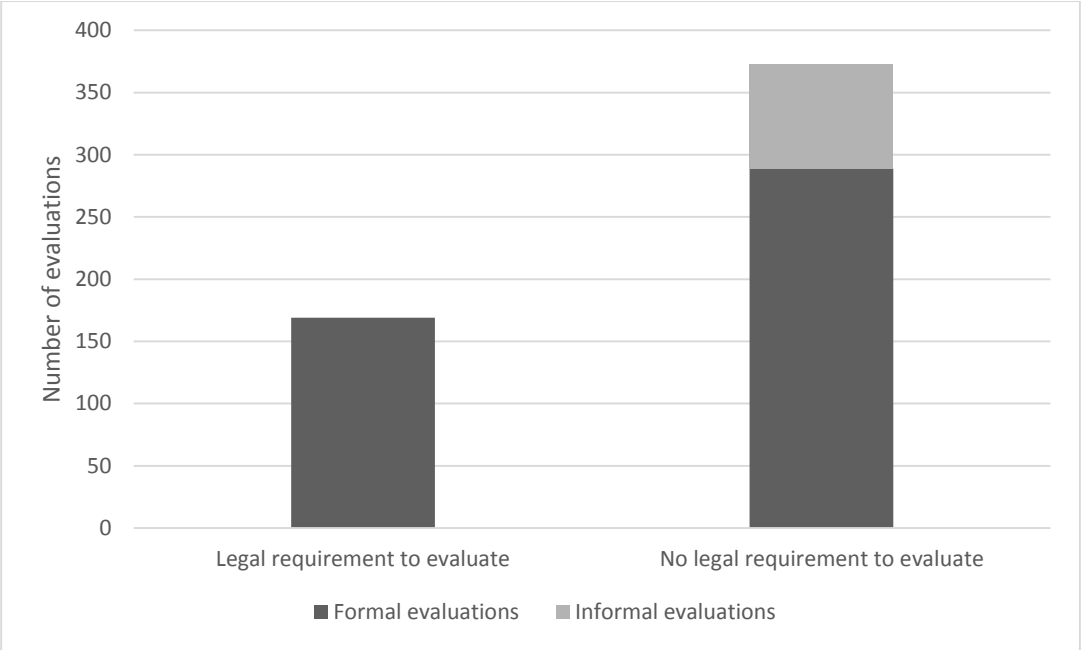


As Chapters 5 and 6 have argued, one way to stimulate policy evaluation is by setting legal evaluation requirements. Legally mandated evaluation may thus be considered least self-organized. Figure 7.9 compares the number of evaluations that have emerged in response to legal requirements to those that have not. It reminds us that there were no informal evaluations that responded to legal requirements (see Chapter 6). Figure 7.9 shows that when looking across the entire database, legal requirements were not the main driver of climate policy evaluation. Other reasons,

[55] Extrapolation for formal evaluations; evaluations whose funders could not be determined are excluded.

such as learning or accountability (see below) account for the majority of evaluations.

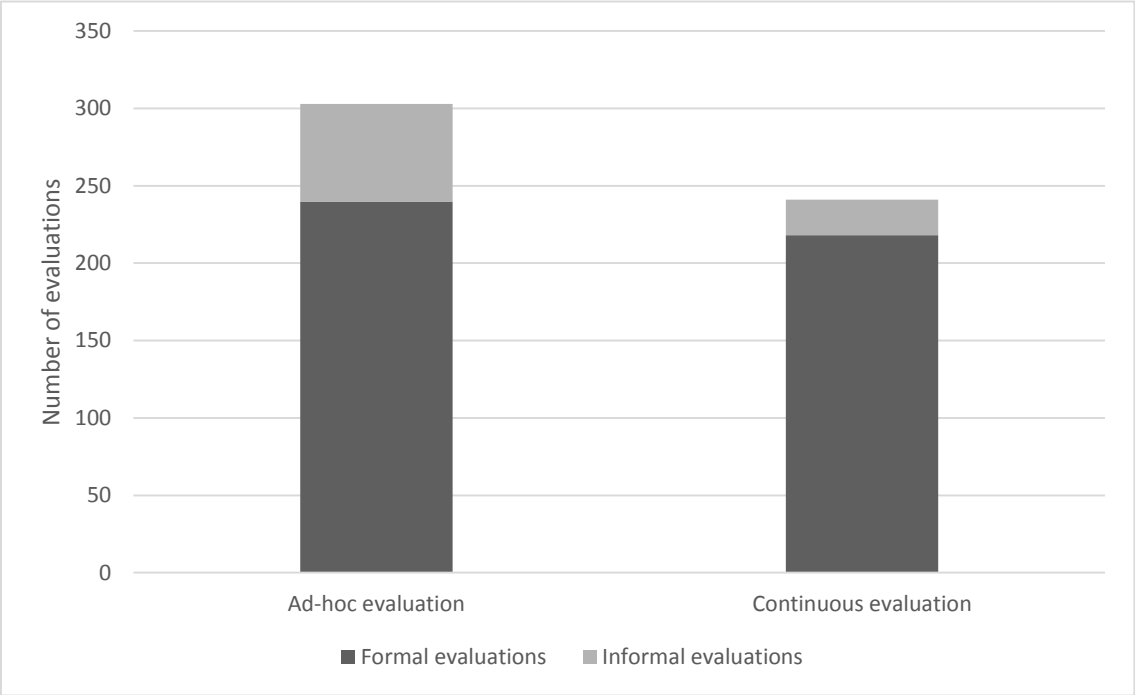
Figure 7.9: Evaluation responding to a legal requirement (N = 542)[56]



Then there is the question of how the formal-informal comparison plays out across ad-hoc evaluations (i.e. evaluations conducted at one point in time) versus continuous evaluations (i.e. whether there were indications that the evaluations were part of larger on-going evaluation exercises). Again Figure 7.10 shows that most continuous evaluations are formal (90.46%), even though there are also some informal continuous evaluations (about one in ten).

[56] Extrapolation for formal evaluations; evaluations whose funders could not be determined are excluded.

Figure 7.10: Ad-hoc versus continuous evaluation (N = 542)[57]



- In sum, looking across the formal-informal comparison reveals that self-organization (the first foundational idea of polycentric governance, see Chapter 2) in climate policy evaluation is rather limited (but the ratio between formal and informal evaluation varies somewhat by governance centre).
- But there has also been growth in self-organized evaluations since 2010, showing that this type of evaluation is nonetheless on the rise.
- Legal requirements do not stimulate any self-organized evaluations, but legal requirements were also not the sole or even the main motivator of formal evaluations.
- Both formal and informal climate policy evaluation tends to happen within individual governance centres, such as at EU level, in Germany and in the UK. The analysis reveals a strong congruence between the location of evaluation funders, the location of the evaluators and the location of the climate policy under evaluation, such that EU level actors tend to mainly fund EU level

[57] Extrapolation for formal evaluations; evaluations whose funders could not be determined are excluded.

evaluators to evaluate EU level climate policy, and so on for Germany and for the UK. Notably, this dynamic is very similar for formal and informal evaluations (see also Chapters 5 and 6).

- Looking beyond one's own governance centre by the means of evaluation is thus an exception rather than a norm (more of which below).
- Finally, the extrapolations calculated in this section revealed that formal actors funded similar numbers of ad-hoc and continuous evaluations, whereas informal evaluations tended to be more ad-hoc.

7.3 Context

The first block of contextual variables concerns those that together form the 'context index' presented in Chapters 5 and 6. Recall that this index includes eight individual variables, namely the time horizon in an evaluation, as well as attention to policy goals, other sectors, unintended policy outcomes, external events, the political environment, geography, and climate science (see Chapters 5 and 6). Figure 7.11 first compares the average composite scores on the context index from formal (N = 84) and informal (N = 84) climate policy evaluations.

Figure 7.11: Context index for formal and informal evaluations (N = 168)

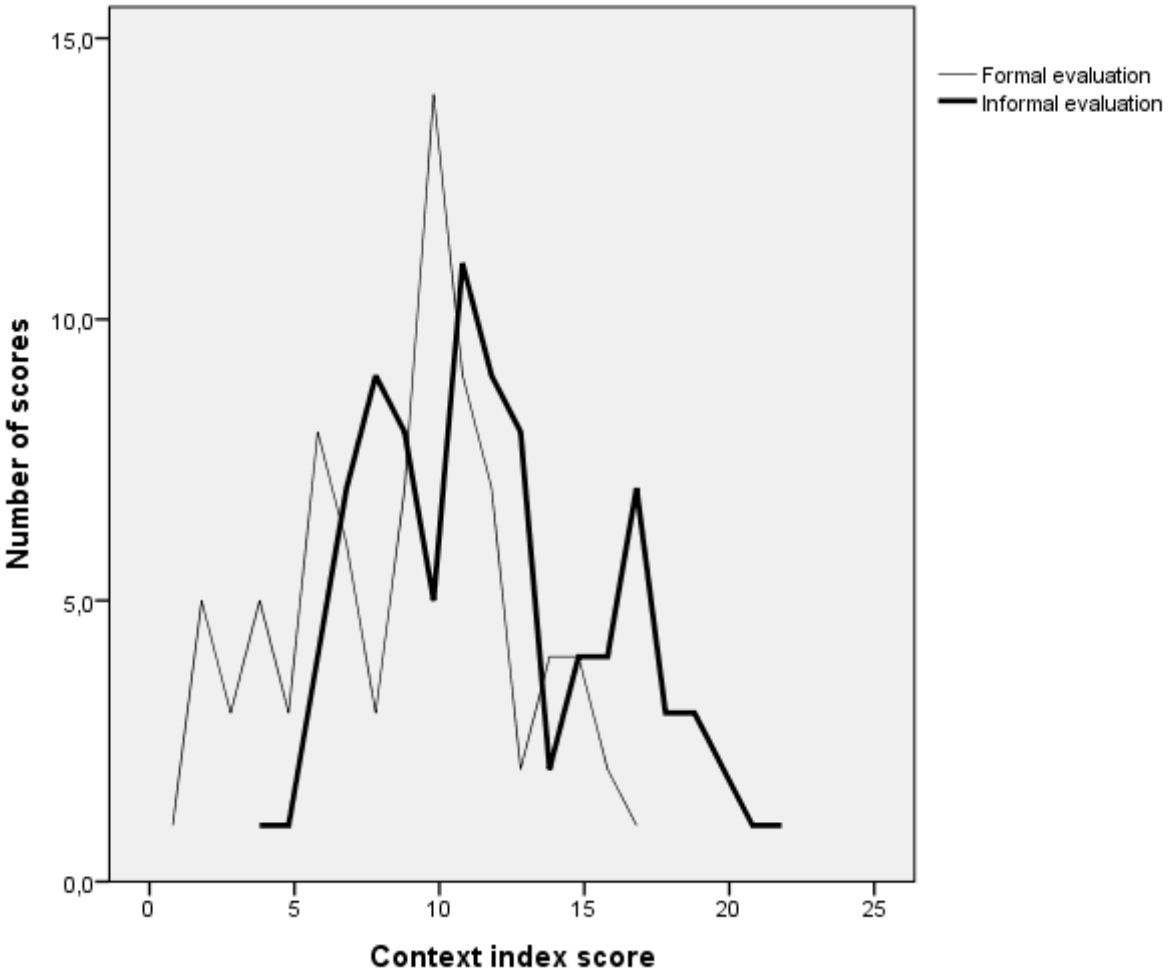
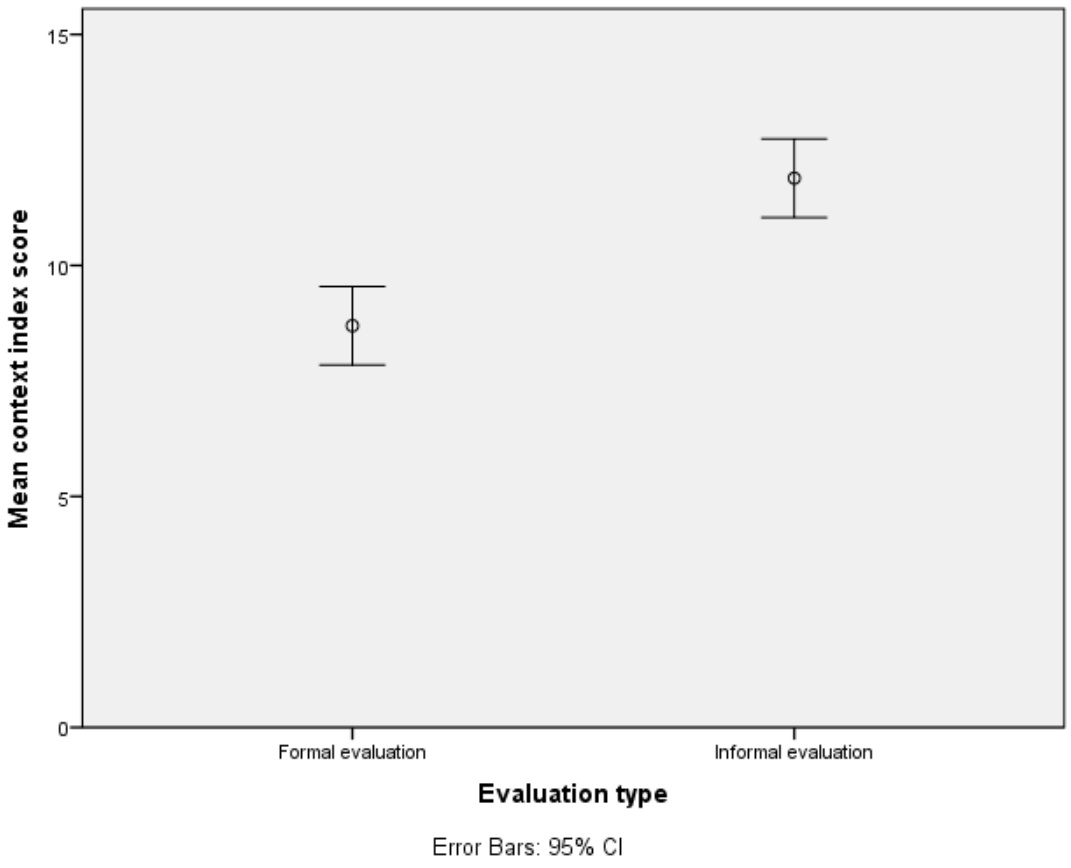


Figure 7.11 shows that there is a difference between formal and informal evaluations on the contextual dimensions. Overall, there are more informal evaluations with a higher context-score (the bold line) than formal evaluations (the thin line). On average, the informal evaluations have a mean score of M = 11.89 (SD = 3.92) while the formal evaluations have a composite mean context score of M = 8.70 (SD = 3.91).[58] An independent samples t-test to compare the means reveals that this difference is also statistically significant with $t(166) = -5.29$, $p < .001$ (two-

[58] There is a lively debate on the extent to which researchers can (or should not) use parametric statistical tests with Likert-type ordinal scales. Following the testing and advice from De Winter and Dodou (2010), who show that in most cases, 5-point ordinal Likert scales approach interval scales in statistical testing, I have used parametric tests with regard to the context index, which contains a mixture of ordinal and interval data, but always highlight that given the ordinal scales, the means comparison is indicative only.

tailed, equal variances assumed). However, it should also be noted that there is still considerable overlap between the two distributions in Figure 7.11. Figure 7.12 however demonstrates the means comparison graphically, showing that the confidence intervals do not overlap, which is an indicator of a high probability of group difference.

Figure 7.12: Average context score for formal and informal evaluations (N = 168)



Overall, these findings demonstrate that informal evaluations on average pay more attention to context according to the dimensions included in the index (see Chapters 5 and 6) than the formal evaluations. But which precise variables drive this difference, or in other words, on which individual variables are the differences between formal and informal evaluations the greatest? In order to test the difference among individual variables, I conducted a range of individual t-tests, comparing all the variables in a pairwise fashion. Table 7.1 presents the results from this statistical comparison.

The table reveals that key differences on three individual variables mainly drive the mean difference between formal and informal evaluations: attention to other sectors (M[formal] = 0.89, M[informal] = 1.65), unintended policy outcomes (M[formal] = 0.76, M[informal] = 1.76), and the political environment (M[formal] = 0.75, M[informal] = 1.64). In each case, informal evaluations had statistically significant higher scores ($p \leq 0.01$) than formal evaluations. The other variables did not return statistically significant differences by evaluation funder type. These findings confirm what polycentric governance scholars would argue (see Chapter 2): informal evaluations generally explore aspects that tend to be highly political and thus potentially uncomfortable for state-related actors to a greater extent than formal evaluations. These aspects include unintended policy outcomes (whose presence may suggest the inability of policy-makers to control outcomes), impacts on other sectors (again something that is difficult to predict and showcases limited control) and, finally, the political environment.

Table 7.1: Context in formal and informal evaluations (N = 168)[59]

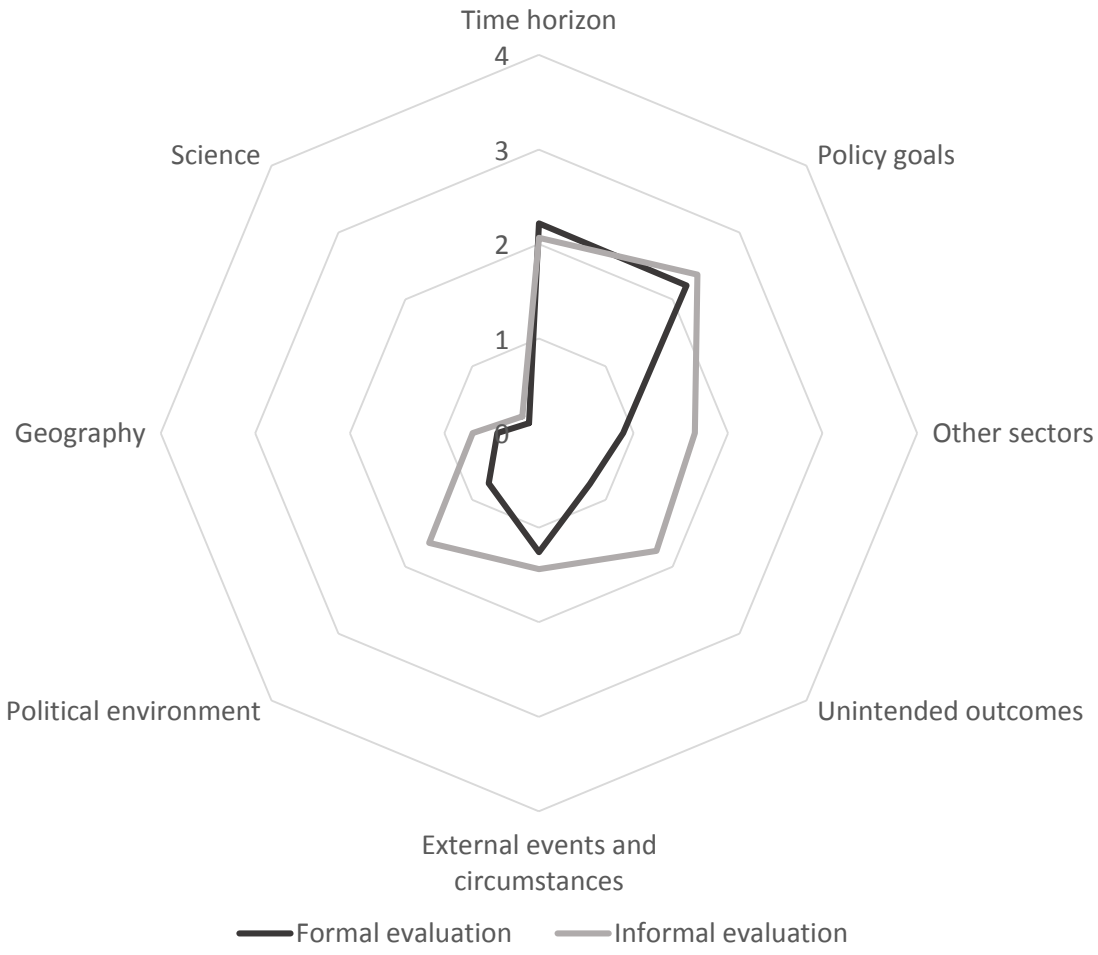
Variables	*M* (formal)	*SD* (formal)	*M* (Informal)	*SD* (informal)	*t*	*df*
Time horizon	3.77	1.68	3.58	1.86	.70	166
Policy goals	2.21	.96	2.37	1.14	-.95	166
Other sectors	.89	1.08	1.65	1.29	-4.15**	160.64†
Unintended outcomes	.76	1.15	1.76	1.43	-5.00**	158.64†
External events	1.26	1.15	1.44	1.27	-.95	166
Political environment	.75	.86	1.64	1.26	-5.37**	146.93†
Geography	.44	.72	.70	1.14	-1.78	139.93†
Science	.15	.43	.25	.66	-1.12	142.191†

* $p \leq .05$
** $p \leq .01$
† Equality of variances not assumed (significant Levene's test).

[59] For variables on an ordinal scale, the means should be considered indicatively only.

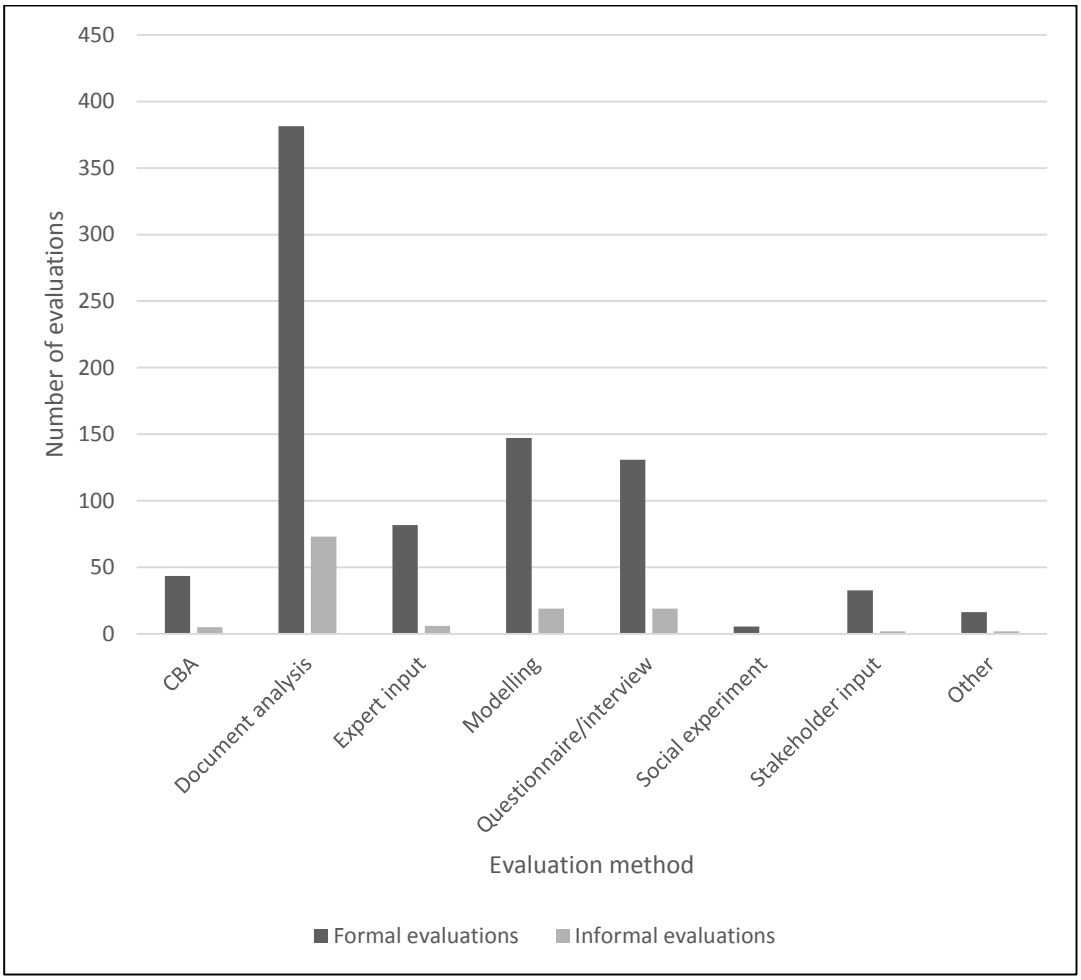
Analogous to Chapters 5 and 6, Figure 7.13 presents the individual contextual variables on a spider diagram, whose vertical axis features the familiar scale (0-4) and where each contextual variable is plotted on each of the rays of the diagram. Reverting back to the argument above, it is immediately visible in Figure 7.13 how informal evaluations contextualize more on some dimensions. However, it also shows common gaps (i.e. dimensions where neither formal nor informal evaluations contextualize much), such as science or geography. It should also be noted that neither formal nor informal evaluations extend, on average, to a 3 on the scale (which corresponds with good attention to the respective contextual variable).

Figure 7.13: Contextual variables in formal and informal evaluations (N = 168)



As Chapters 5 and 6 have highlighted, the factors analysed above are not the only ones that matter when assessing attention to context in evaluation. An especially important additional perspective concerns the method(s) that an evaluation uses. First, there is the *type* of methods. Figure 7.14 demonstrates that while almost all formal and informal evaluations use document analysis as a method, there are significantly more formal evaluations that use expert input, questionnaires and CBA than informal evaluations. Thus, interestingly, formal evaluations also cover the more participatory evaluation methods (such as direct stakeholder input or questionnaires) rather well.

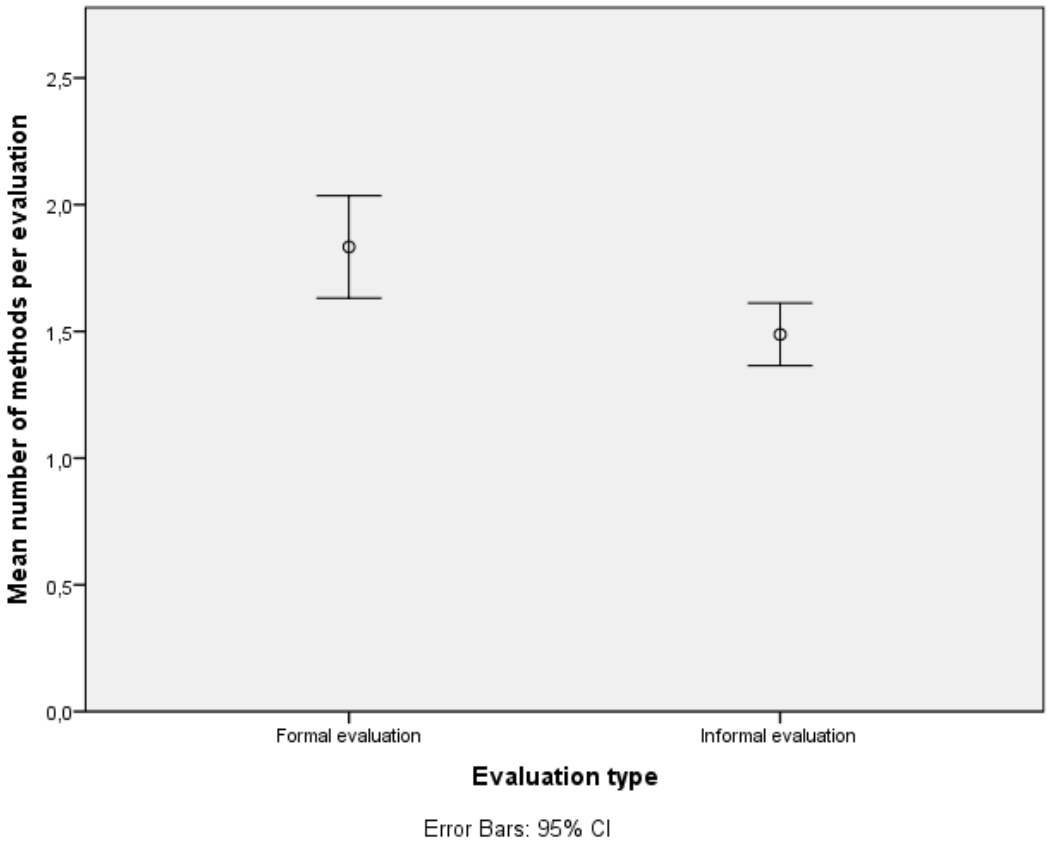
Figure 7.14: Types of methods in formal and informal evaluations (N = 542)[60]



[60]Multiple mentions possible; extrapolation for formal evaluations; evaluations whose funders could not be determined are excluded.

Then there is the number of evaluation methods used in evaluations. A higher number of methods could indicate greater efforts to disentangle various policy effects, as each method has unique strengths and weaknesses, which translate into analytical strength in some areas, but also potential blind spots in others (see Chapter 2). Taking the data from Chapters 5 and 6 together (N = 168), there is a difference between formal and informal evaluations, such that on average, formal evaluations use a greater number of methods (M = 1.83, SD = 0.93) than informal evaluations (M = 1.49, SD = 0.57). A t-test to compare the two means shows that this difference is also statistically significant with $t(138) = 2.90$, $p < 0.01$ (two-tailed, equal variances not assumed due to a significant Levene's test). Figure 7.15 shows the means comparison and again highlights that there is very little overlap in the error bars, which is another indicator that the difference is significant.

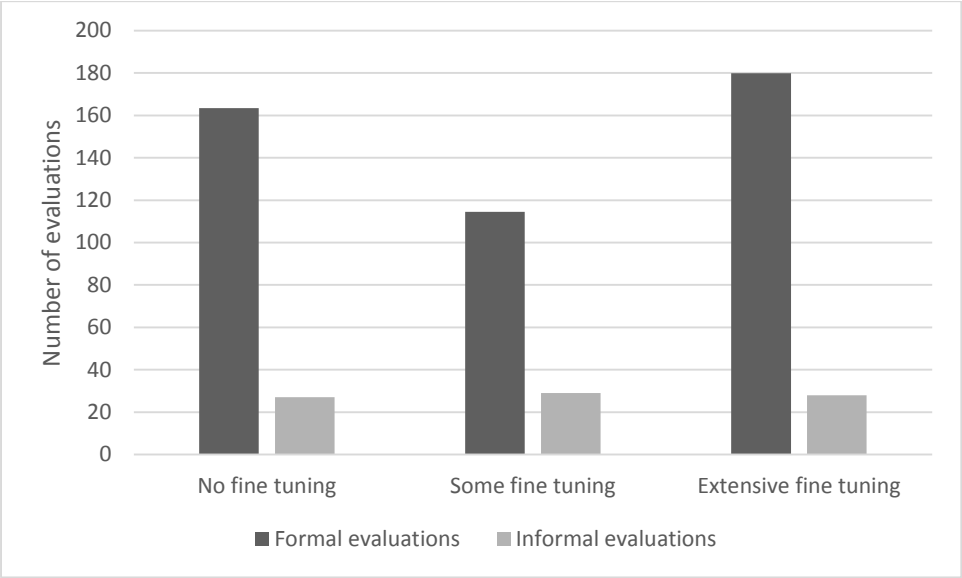
Figure 7.15: Average number of methods used in formal and informal evaluations (N = 168)



In line with the discussion in Chapter 2, utilizing a greater number of methods could also point to greater resource availability among formal evaluation funders in order to use various methods. However, one should also note that the formal evaluations only use on average about 'half' a method more than the informal evaluations. Furthermore, the range of observed scores is higher for formal evaluations (maximum = 5 methods) as compared to the informal evaluations (maximum score = 3 methods).

But in addition to the number of methods, Chapters 5 and 6 revealed that the calibration (or fine-tuning) of the methods matters because this affects an evaluation's ability to cater to and perhaps better detect contextual policy effects. However, the level of methodological calibration between formal and informal evaluations is very similar. As Figure 7.16 demonstrates, although the overall difference in numbers of formal and informal evaluation is of course evident (see also above), the relative differences are small. Fully 35.71% of the formal evaluations exhibited no methodological fine-tuning (with 32.14% of the informal evaluations), while 25% of the formal evaluations contained some fine tuning (34.52% for informal evaluation), and 39.29% of the formal evaluations contained extensive fine tuning (with 33.33% for the informal evaluations). While there is thus some fluctuation, there is no clear trend in either direction. These findings thus suggest that as far as the calibration of evaluation methods is concerned, it does not appear to matter whether the funders are formal or informal – both are equally low (recall that a '1' on the respective scale meant 'some fine-tuning' – see Appendix 3).

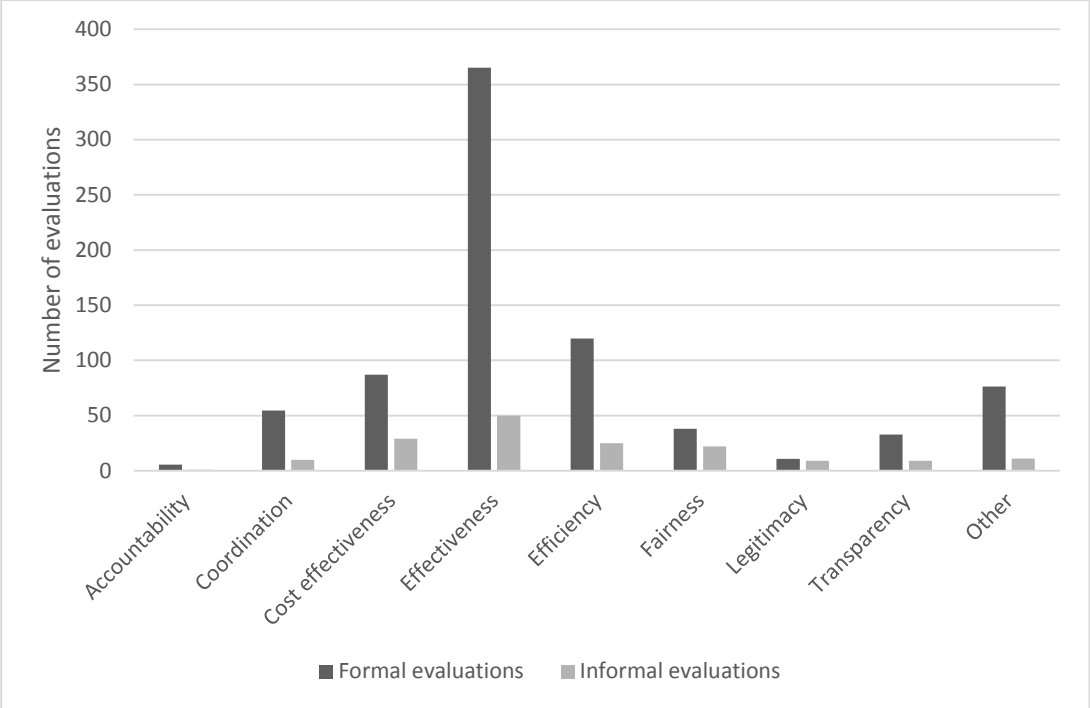
Figure 7.16: Methodological calibration in formal and informal evaluations (N = 542)[61]



Analogous to the discussion of methods in evaluation, Figure 7.17 presents the *types* of evaluation criteria, which formal and informal evaluations used in order to assess climate policies. It reveals that while there was a greater number of formal evaluations than informal evaluations in every category, informal evaluations demonstrated a disproportionately greater propensity to assess whether the climate policy was fair and legitimate. Informal evaluations thus focused especially on social criteria that were much less applied in the formal evaluations. However, virtually no evaluation focused on accountability, and the number of evaluations that evaluated the legitimacy of the climate policy in question was very low for both formal (an estimated 10 evaluations) and informal evaluations (9).

[61] Extrapolation for formal evaluations; evaluations whose funders could not be determined are excluded.

Figure 7.17: Types of criteria in formal and informal evaluations (N = 542)[62]



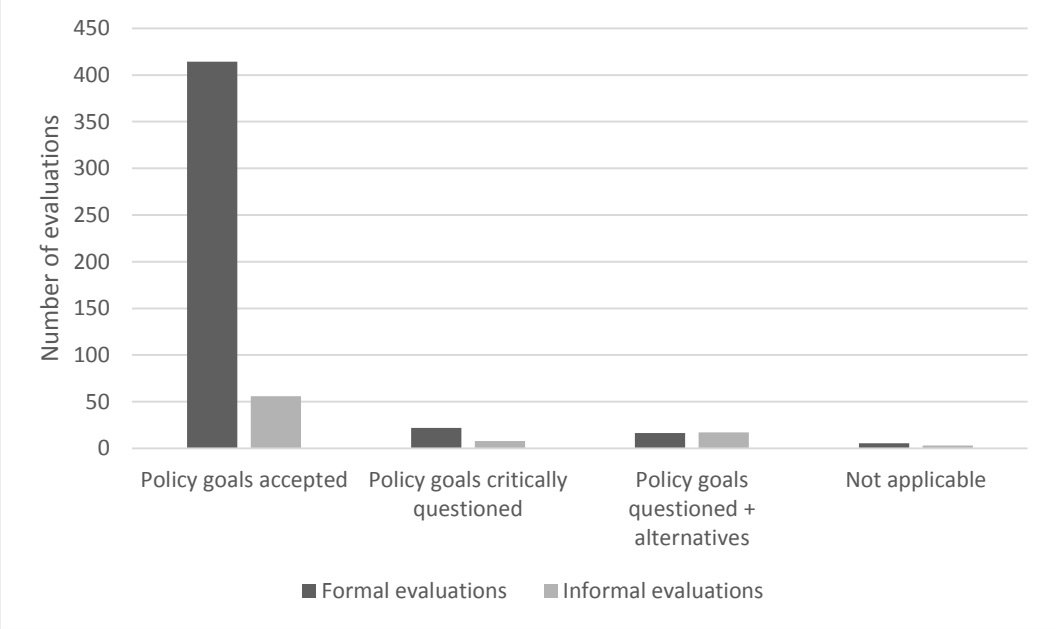
A similar logic also applies to the number of criteria that the evaluations apply. A greater number of evaluation criteria points to efforts to capture more contextual effects (see Chapter 2). Here, formal evaluations use on average somewhat fewer criteria (M = 1.73, SD = .87) than informal evaluations (M = 1.98, SD = 1.09). A statistical comparison focusing on the sample of the formal evaluations and all informal evaluations returns a statistically insignificant difference between the means from formal and informal evaluations where $t(166) = -1.65$, *ns* (two-tailed, equal variances assumed). Note that in each case, the maximum number of criteria used in an evaluation was 5.

As Chapters 5 and 6 have already explained, reflexivity – or the extent to which evaluations critically question extant policy goals and/or offer alternatives – is another important way to account for policy context in evaluations. Analysing reflexivity offers an insight into the extent to which extant policy targets (still) fit the context which prevailed at the time when the targets were adopted. If the context has changed, older targets may no longer be deemed relevant or appropriate at the time

[62] Multiple mentions possible; extrapolation for formal evaluations; evaluations whose funders could not be determined are excluded.

of evaluation. Figure 7.18 demonstrates that formal evaluation are, proportionately speaking, less reflexive than informal evaluations. While the vast majority of formal evaluations exhibited no reflexivity (fully 90.48% accepted extant policy goals as given), this is only true for 66.67% of the informal evaluations. Only 4.76% of the formal evaluations critically questioned the policy goals, but 9.52% of the informal evaluations did so. Importantly, while only 3.57% of the formal evaluations critically questioned policy goals and proposed alternatives, 20.24% of the informal evaluations engaged in this way. In sum, informal evaluations are considerably more reflexive than formal evaluations. This, again, confirms the expectation of polycentric governance and evaluation theory that informal evaluations may be more reflexive.

Figure 7.18: Reflexivity in formal and informal evaluations (N = 542)[63]



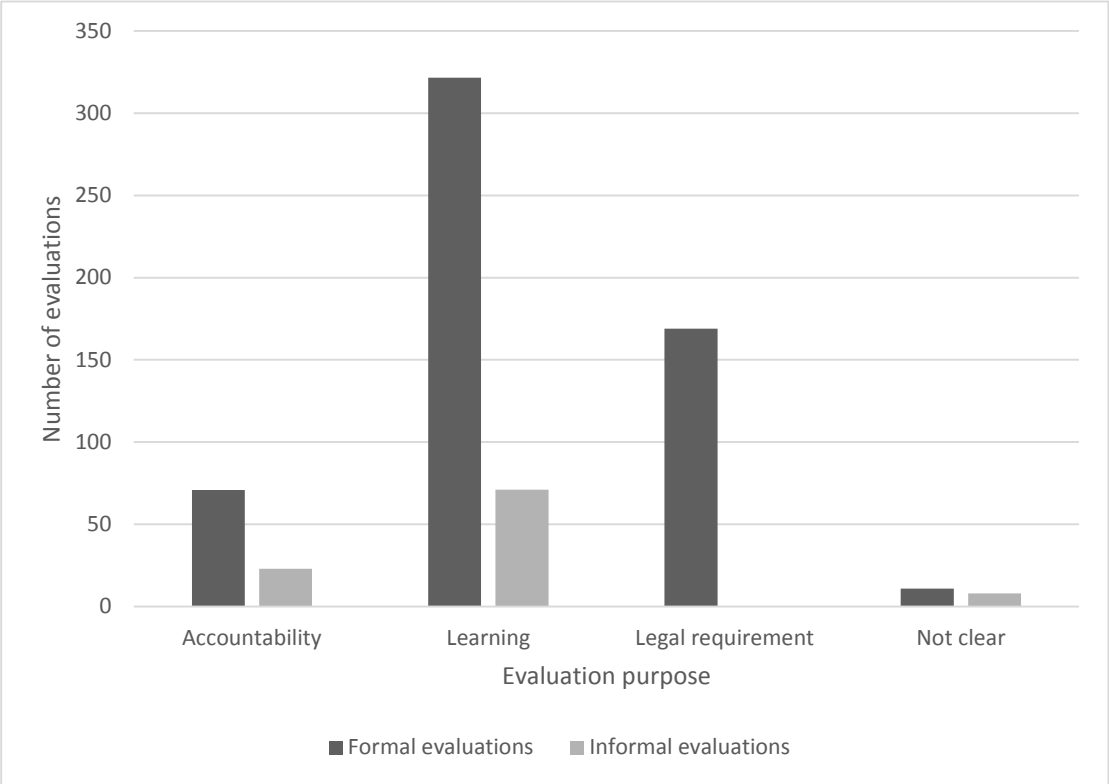
[63] Extrapolation for formal evaluations; evaluations whose funders could not be determined are excluded.

- Looking across all the comparisons reveals that, on the whole, informal evaluations tend to pay more attention to contextual variables, and they are more reflexive than formal evaluations. Differences on references to the political environment, unintended side effects, as well as policy effects in other sectors drove the difference between formal and informal evaluations on the context index explained above and in the previous chapters.
- There is no difference between formal and informal evaluations regarding methodological tailoring, but formal evaluations used a significantly greater number of methods than informal evaluations. This may point to greater levels of available resources and a willingness to use a greater number of methods among formal evaluation funders; by the same token, informal evaluations evidently explore other aspects, including more political ones, in greater depth (for a discussion of the theoretical implications of these findings, see Chapter 8).

7.4 Interaction

The third foundational idea in polycentric governance theory is the extent to which evaluations support interactions between different governance centres. Figure 7.19 presents a comparison between the formal and informal climate policy evaluations with regard to the declared evaluation purpose (numbers for the formal evaluations have been extrapolated from the sample using the methodology explained in the above sections). Recall that the evaluation purpose may significantly influence evaluation use later on. While Figure 7.19 reflects the fact that the number of formal evaluations is significantly greater than the number of informal evaluations (see above), it is also notable how the proportions of the number of mentions of the different purposes appear to mirror each other across the two types. For both formal and informal evaluation, learning and improvement is the most frequently stated purpose, followed by accountability. Only formal actors funded climate policy evaluation in response to legal requirements. The generally strong focus on learning may be one indication that the evaluations analysed here endeavour to provide lessons and stimulate improvement – either for the policies they evaluate or even for other governance centres.

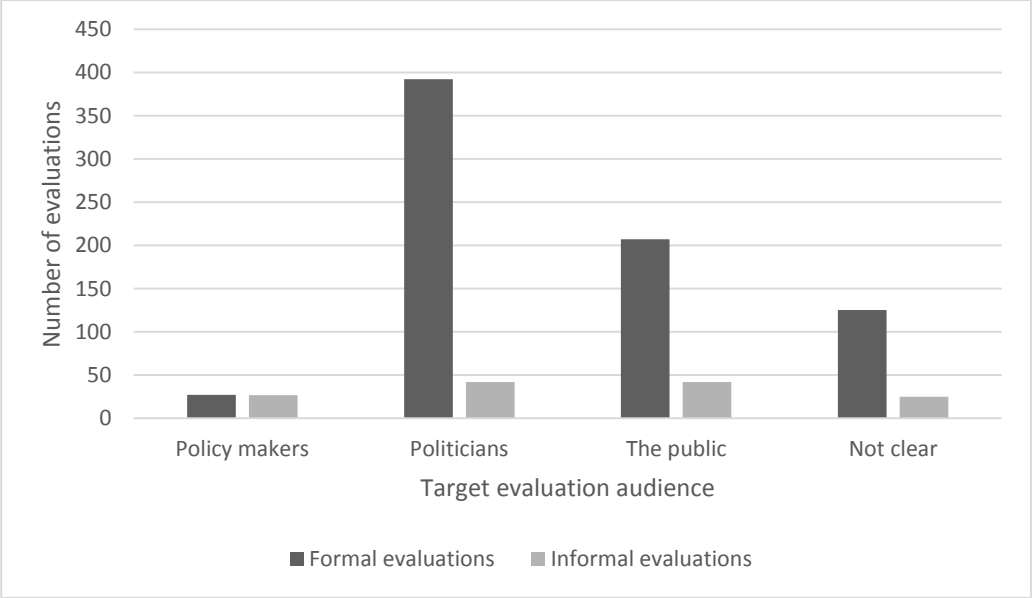
Figure 7.19: Evaluation purpose (original N = 542)[64]



As Chapters 5 and 6 highlighted, in addition to the stated evaluation purpose, the extent to which an evaluation may support interactions between governance centres also depends on the target audience of the evaluation. Therefore, Figure 7.20 compares the target audience for formal and informal evaluations; as before, the numbers of formal evaluations have been extrapolated. It reveals that the differences between the different types of audiences are much larger among formal evaluations than among the informal ones (i.e. comparing the height of the bars with the same hue). Whereas the informal evaluations stated that they are geared towards policy-makers and politicians in equal numbers, the target audience of formal evaluations is much more directed towards policy-makers, followed by politicians, and then finally the general public.

[64] Multiple mentions possible; extrapolation for formal evaluations.

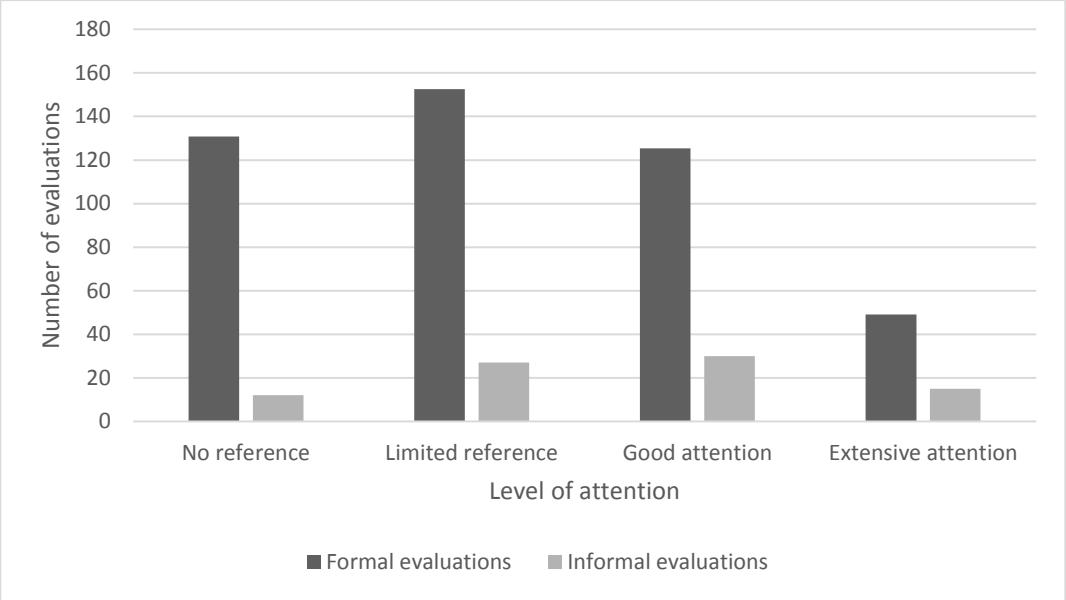
Figure 7.20: Target evaluation audience (original N = 542)[65]



But perhaps one of the most straightforward ways of expressing interactions directly is through the extent to which formal and informal evaluations draw on insights from other evaluations focusing on the same centre. Figure 7.21 presents the respective data, and shows that proportionately speaking, informal evaluations show a stronger tendency to draw on insights from evaluations of the same centres than form al evaluations. More specifically, 28.57% of the formal evaluations made no reference to insights from the same centre, with the analogous number of informal evaluations being 14.29%. While the numbers for limited reference to insights from the same centre are fairly similar (33.33% for formal and 32.14% for informal evaluations), 27.38% of the formal evaluations paid good attention to insights from the same centre (with 35.71% for informal evaluations). Most importantly, only 10.71% of the formal evaluations paid extensive attention to insights from the same centre, with fully 17.86% of the informal evaluations doing so.

[65] Multiple mentions possible; extrapolation for formal evaluations.

Figure 7.21: Level of attention to evaluations of the same centre (original N = 542)[66]

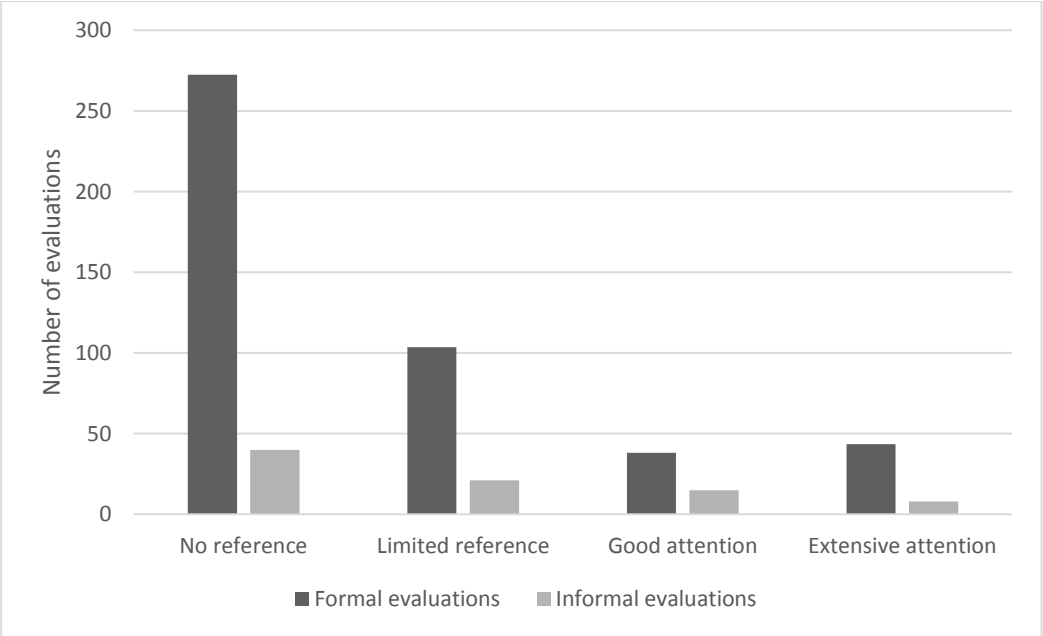


However, this difference does not emerge as strongly from the data for an analogous, and yet substantially different, way of looking at interactions between governance centres; namely, the extent to which climate policy evaluations draw on insights from evaluations of other governance centres. Across the scale, the data are fairly similar: for no reference to insights from other governance centres (59.52% of the formal evaluations and 47.62% of the informal evaluations); for limited reference (22.62% of the formal evaluations and 25.00% of the informal evaluations); there is a difference for 'good attention for insights from other governance centres' with only 8.33% of the formal evaluations exhibiting this characteristic, but 17.86% of the informal evaluations; and finally the same proportion of formal and informal evaluations paying extensive attention to insights from other governance centres (both 9.52%). Figure 7.22 represents the respective data. Overall, neither formal nor informal evaluations engaged with insights from other governance centres in depth. However, as the more detailed results in Chapters 5 and 6 demonstrate, there is a small number of outliers at the top of this spectrum that very much engaged with

[66] Extrapolation for formal evaluations.

insights from other governance centres, and informal evaluations have a slightly greater propensity in this regard.

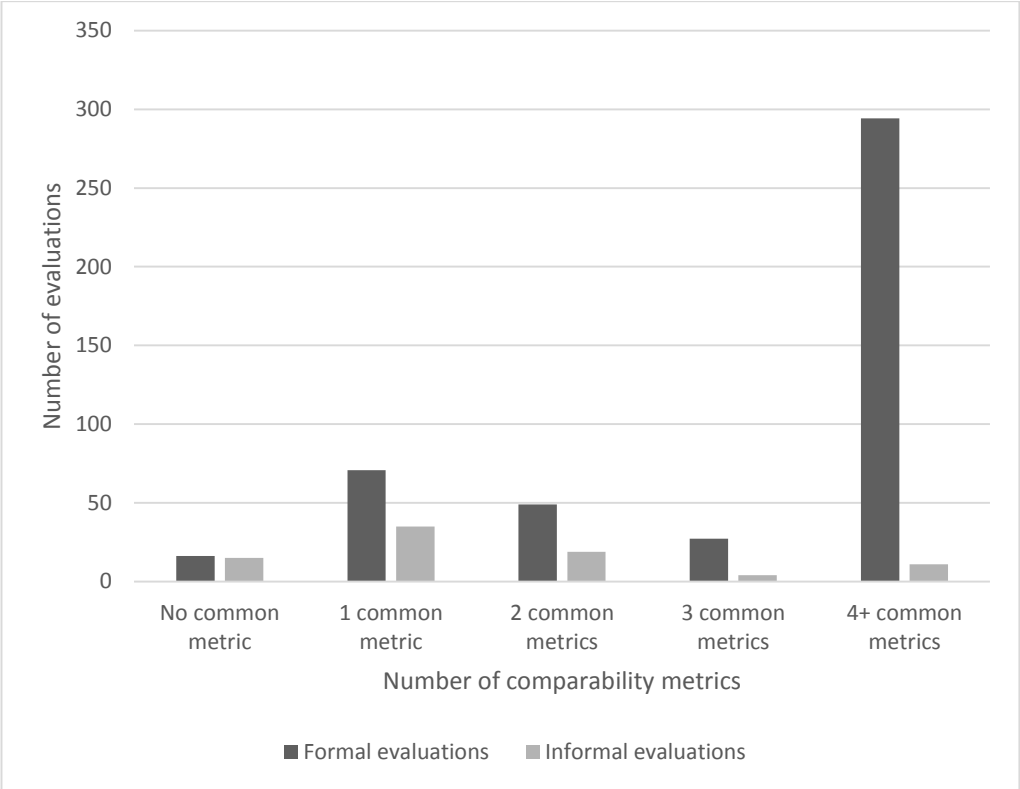
Figure 7.22: Level of attention to evaluations of other centres (original N = 542)[67]



Quantitative comparability metrics are another important way in which evaluation may foster interactions between governance centres by allowing them to compare their experiences. Figure 7.23 reveals that across the whole database, there are many more formal evaluations that contain four or more quantitative comparability metrics than informal evaluations. Furthermore, the share of formal evaluations that use four or more metrics (64.29%) is much higher than the share of informal evaluations that use four or more quantitative metrics (13.10%). Formal evaluations evidently quantify their findings more.

[67] Extrapolation for formal evaluations.

Figure 7.23: Number of comparability metrics in formal and informal evaluation (N = 542)[68]

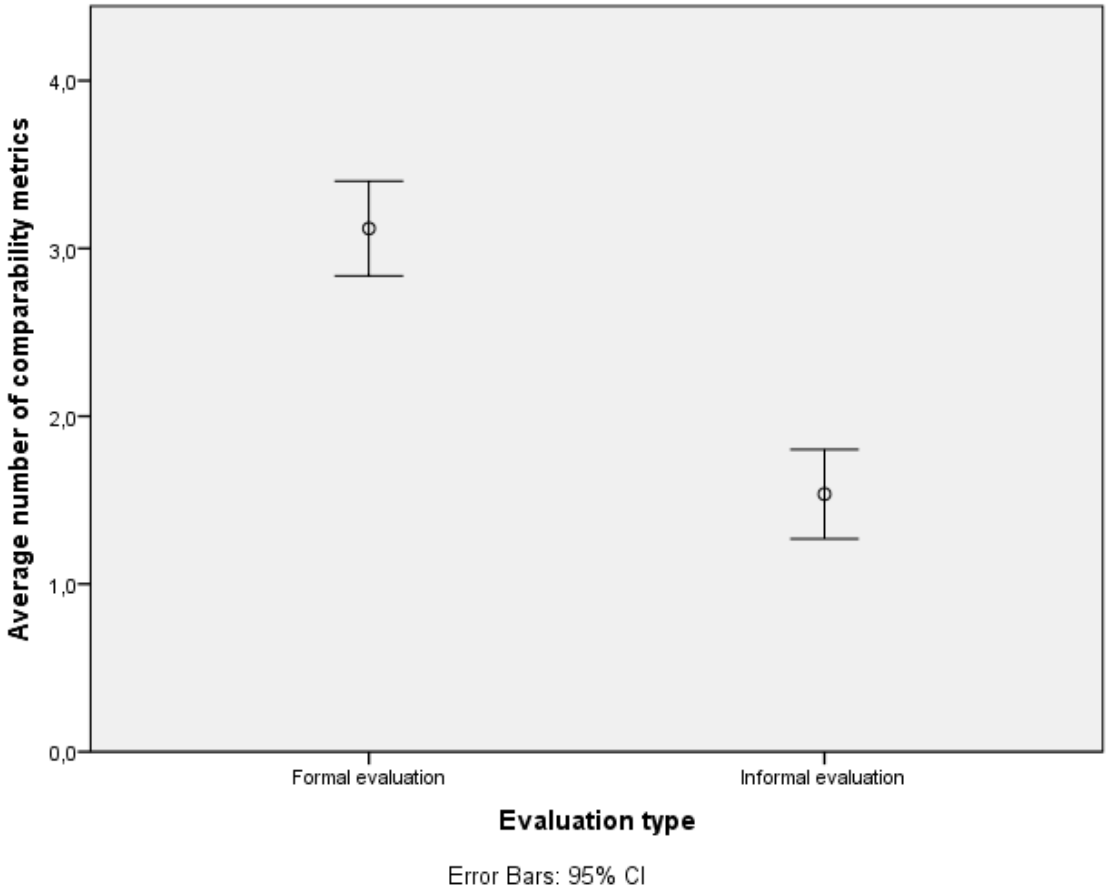


A comparison of the means between the sample of formal evaluations (M = 3.12, SD = 1.30) and the informal evaluations (M = 1.54, SD = 1.23)[69] reveals that this this difference is also highly statistically significant with $t(166) = 8.11$, $p < 0.01$ (two-tailed, equal variances assumed). Figure 7.24 presents this mean difference graphically. No overlap between the error bars indicates high probability of group difference.

[68] Extrapolation for formal evaluations.

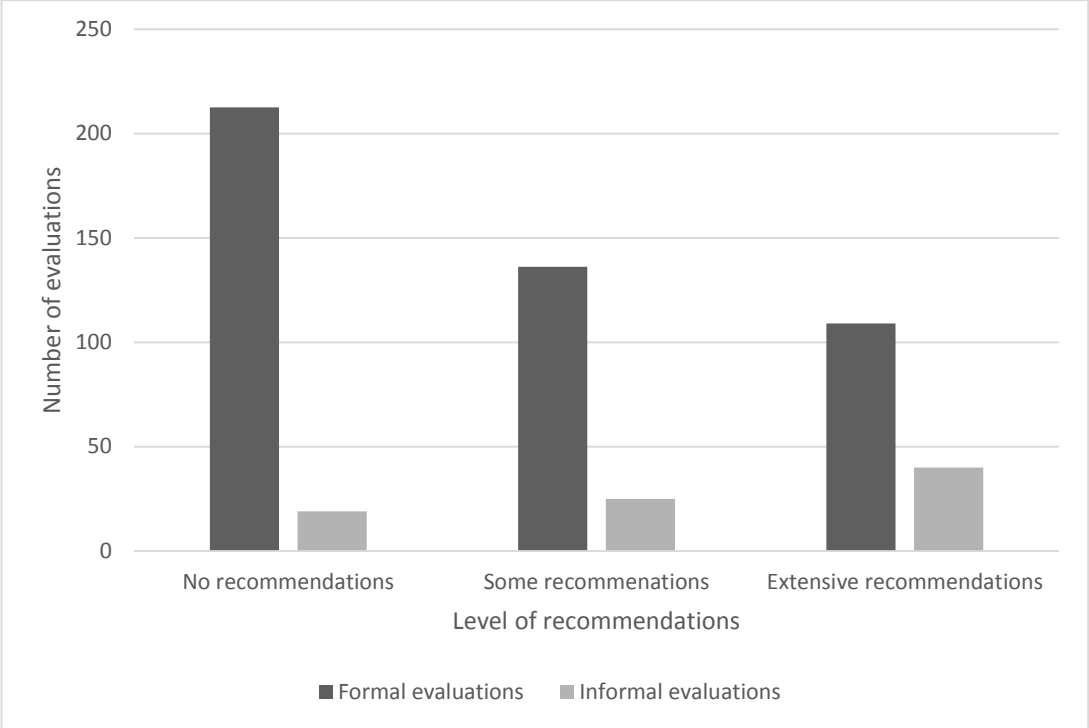
[69] 4+ has been counted as "4"; thus, this difference is likely to be an under-estimation.

Figure 7.24: Average number of comparability metrics (N = 168)



As Chapters 5 and 6 have argued, policy recommendations are another way in which insights about climate policy effects may in principle travel from one governance centre to another. Figure 7.25 reveals a trend where the greatest number of formal evaluations does not contain recommendations (46.43%), and the higher the level of recommendations, the fewer formal evaluations there are. For informal evaluations, the inverse appears to be true. The biggest category here is extensive recommendations (47.62%), with fewer evaluations that contain only ‘some recommendations’ (29.76%) or even ‘no recommendations (22.62%).

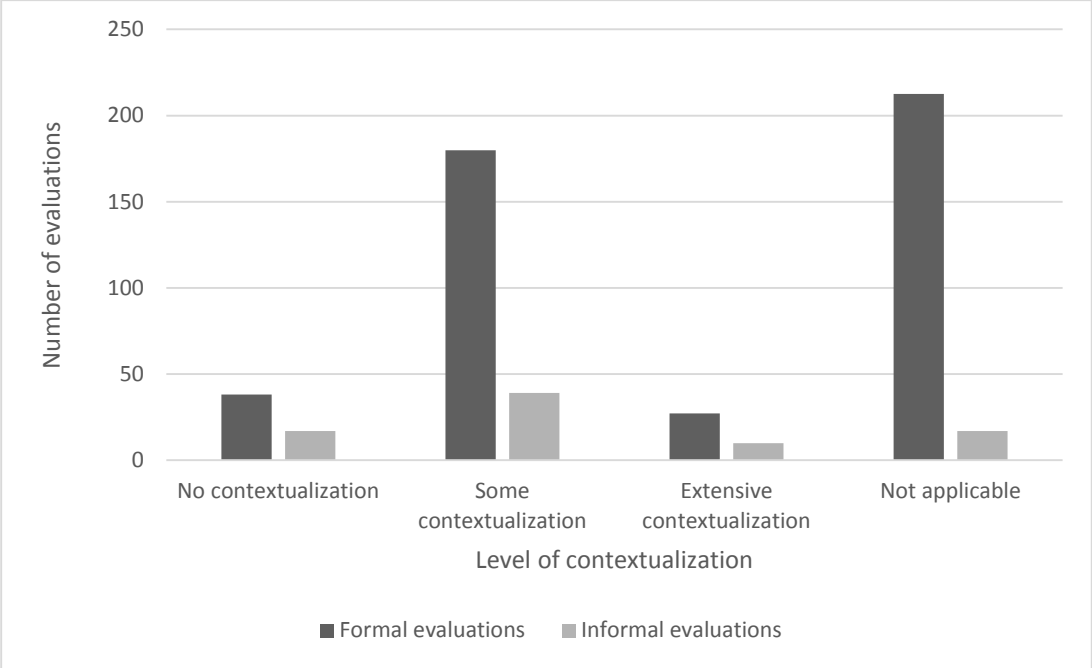
Figure 7.25: Evaluation by level of recommendation (N = 542)[70]



Following the theoretical arguments about the importance of context in polycentric governance (Chapter 2), Chapters 5 and 6 present data on the level of contextual information contained in recommendations in the evaluations. Figure 7.26 compares the level of contextualization of the policy recommendations in formal and informal evaluations. It shows that only 5.95% of the formal evaluations contextualized their recommendations extensively (compared to 12.04% for the informal evaluations); 39.29% of the formal provided some contextualization of their recommendations (compared to 46.99% for informal evaluations) and, finally, 8.33% of the formal evaluations did not contextualize the recommendations at all (with 20.48% of the informal evaluations not contextualizing their recommendations at all). Note also the considerable "not applicable" category, which means that the evaluation did not provide any recommendations in the first place. Taken together, neither formal nor informal evaluations contextualize their recommendations much, but informal evaluations show a slightly greater tendency to do so.

[70]Extrapolation for formal evaluations.

Figure 7.26: Contextualization in policy recommendations (N = 542)[71]

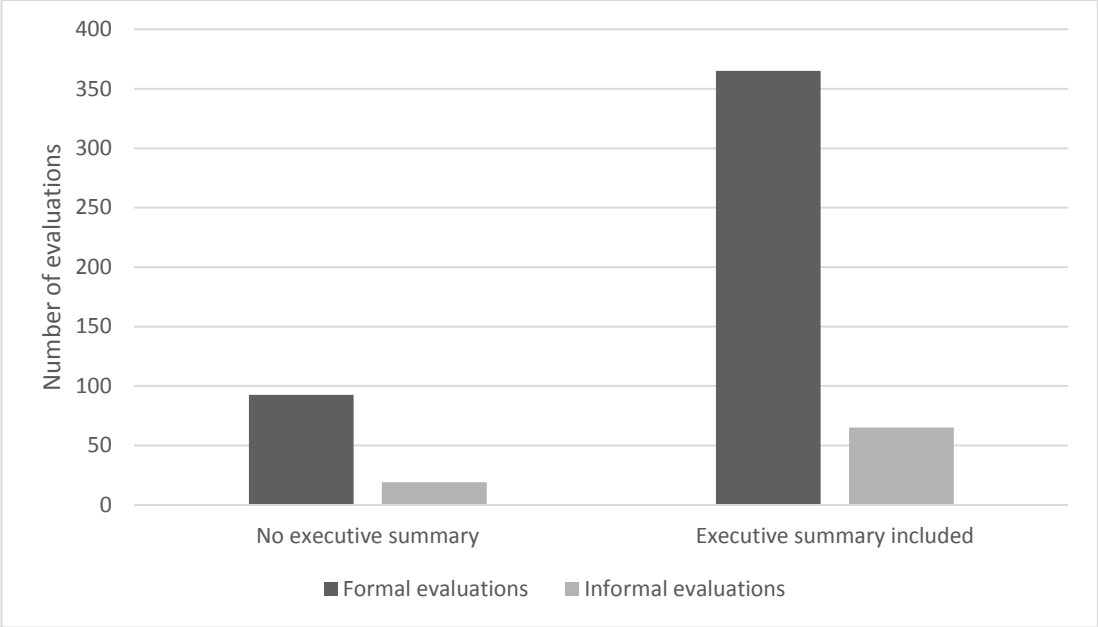


Taking the results of Figure 7.25 and Figure 7.26 together reveals that while informal evaluations are inclined to present *more* policy recommendations, that difference hardly matters for the level of contextualization of the recommendation. However, it should be noted that the high number of evaluations without recommendations (i.e. 'not applicable') somewhat distorts these summary statistics because it leads to an over-estimation.

An additional variable that may influence the extent to which insights can travel from one governance context to another vis-à-vis evaluation is whether evaluations contain executive summaries that highlight the main findings in an evaluation. For busy policy-makers in fast-moving environments, such summaries may help to quickly understand the main conclusions without having to delve into lengthy reports. Figure 7.27 shows that most formal and informal evaluations included executive summaries. In other words, the nature of the funder does not appear to influence whether or not the evaluation contains an executive summary.

[71]Extrapolation for formal evaluations.

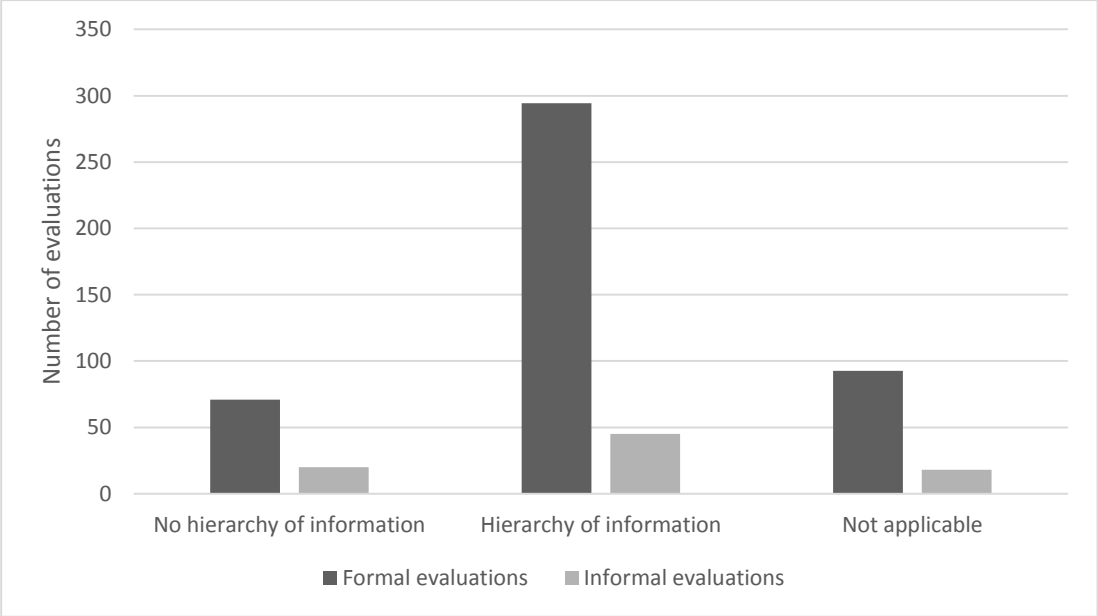
Figure 7.27: Executive summaries in evaluations (N = 542)[72]



As Chapters 5 and 6 have explained, an additional way to make evaluative information from the executive summaries even more accessible to policy-makers and others is to structure the executive summaries. This may be done by providing clear sections, bolding key passages or terms, or through figures and tables that summarize the most relevant information. Figure 7.28 thus compares the extent to which executive summaries are internally structured in formal and informal evaluations. It reveals that there is no large difference on this dimension; a Chi-squared test to assess this difference statistically (based on the sample of formal evaluations and all informal evaluations; total N = 168) reveals no significant difference between the two groups with $X^2(2, N= 168) = 2.32$, *ns.* Taken together, Figure 7.28 demonstrates that the source of evaluation funding does not appear to hang together with the extent to which the information in executive summaries is structured.

[72]Extrapolation for formal evaluations.

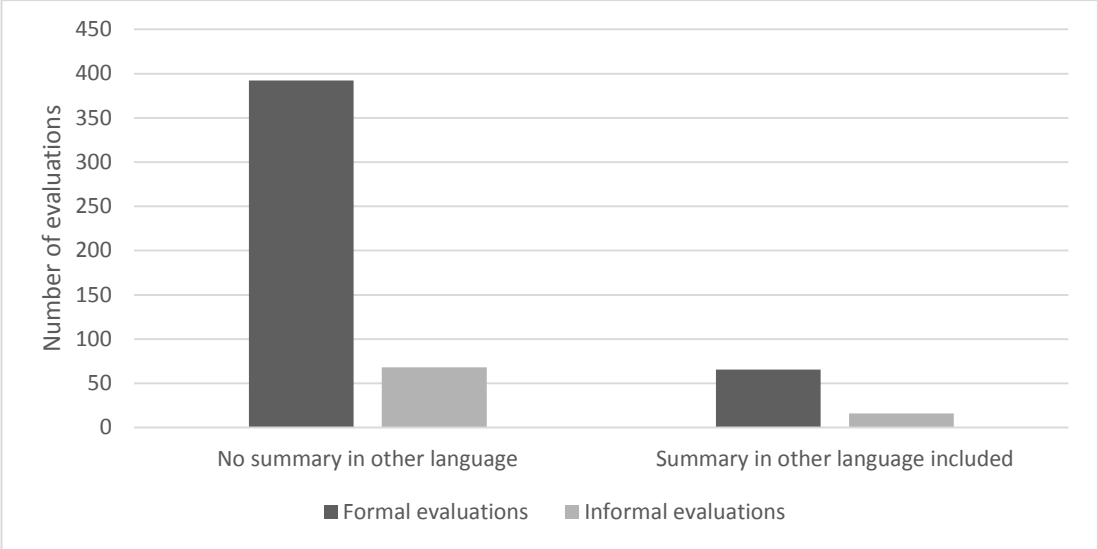
Figure 7.28: Hierarchy of information in executive summaries (N = 542)[73]



Another important dimension of interacting governance centres is linguistic access, a feature that is particularly relevant in multi-lingual environments such as the EU (see Chapter 2). If evaluative insights are to travel from one governance centre to another, they may also have to overcome language barriers. Therefore, Figure 7.29 compares the extent to which formal and informal evaluations contain summaries – or even whole versions – of the evaluation in other languages. It reveals that the source of funding does not appear to influence whether or not evaluations contain summaries in other languages. The number of climate policy evaluations that are available in other languages is severely limited, irrespective of the evaluation funding source.

[73]Extrapolation for formal evaluations.

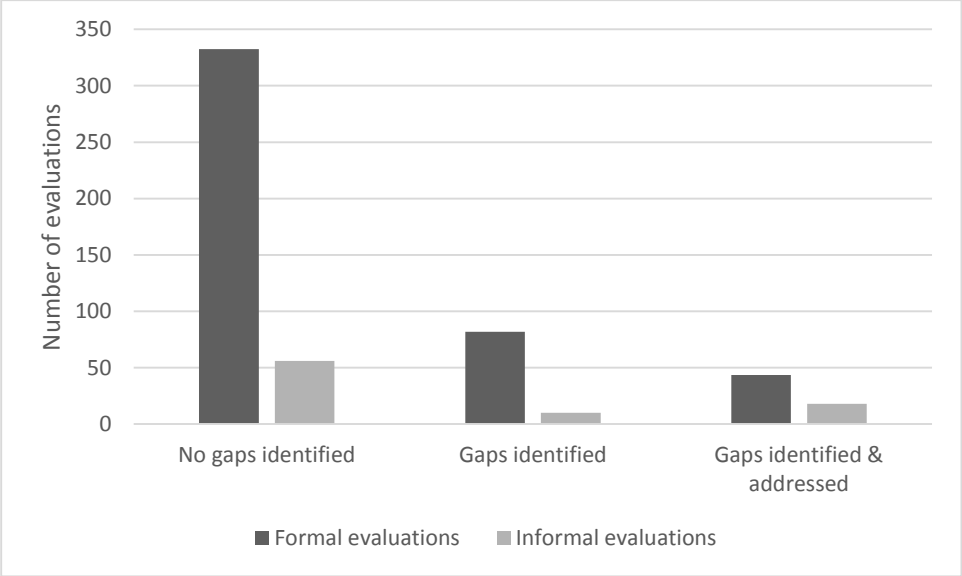
Figure 7.29: Linguistic access in climate policy evaluations (N = 542)[74]



Aside from the rather outward-facing aspects discussed above, there are other aspects internal to the evaluation that indicate interactions across different governance centres. One important aspect is the extent to which informal evaluations aim to fill any gaps left by formal evaluations, and vice versa. For example, this could include an informal evaluation arguing that formal, state-funded data and/or analyses are limited, and that it therefore collects its own data and conducts its own analysis. Figure 7.30 thus compares the extent to which formal and informal evaluation discuss and action this type of interaction. It shows that, across the formal and informal evaluations, the difference is rather limited, even though there is – in light of the overall proportions – a somewhat greater share of informal evaluations (21.43% in comparison to only 9.52% of formal evaluations) that both identify and address gaps, than formal evaluations. Neither formal nor informal evaluations are thus particularly geared towards spotting the gaps left by others and addressing them.

[74]Extrapolation for formal evaluations.

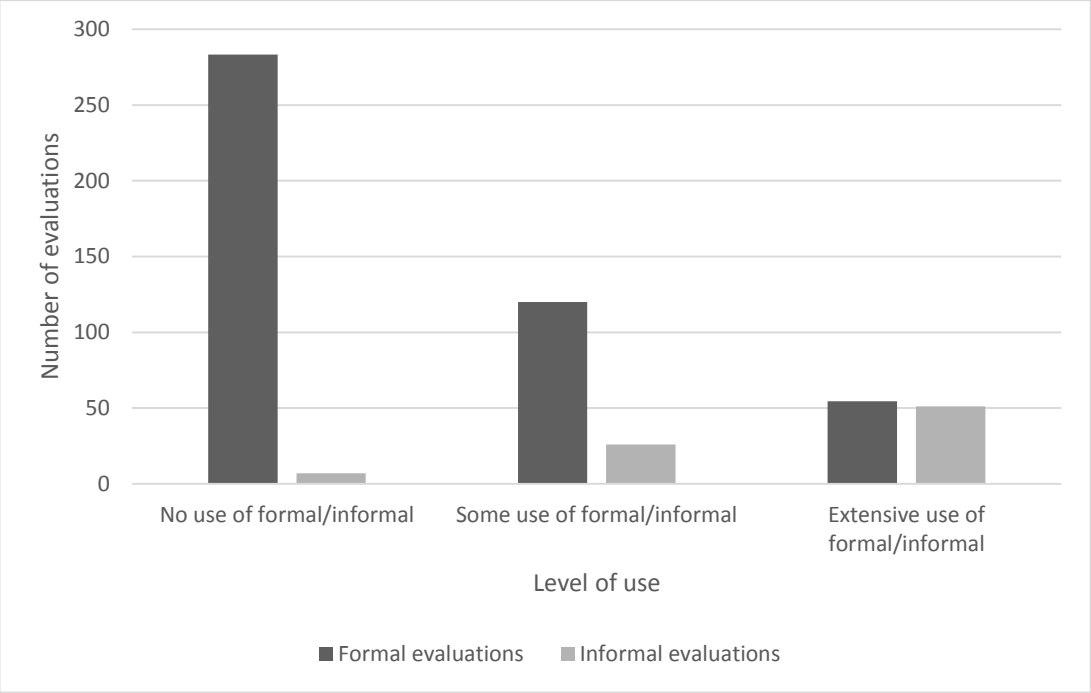
Figure 7.30: Gap filling by formal and informal evaluations (N = 542)[75]



A related aspect concerns the extent to which formal and informal evaluations draw on the findings of other evaluation types. For example, this includes whether informal evaluations draw on insights or even use data from formal evaluations. Figure 7.31 compares the extent to which this happens. It shows that there is indeed a significant difference between the groups; while the majority (61.90%) of formal evaluations makes no use of insights or data from informal evaluations, at the other end of the spectrum, the majority of informal evaluations (60.71%) uses insights or data from formal evaluations extensively compared to only 11.90% of the formal evaluations. Informal evaluations thus use more data and insights from formal evaluations than vice versa.

[75]Extrapolation for formal evaluations.

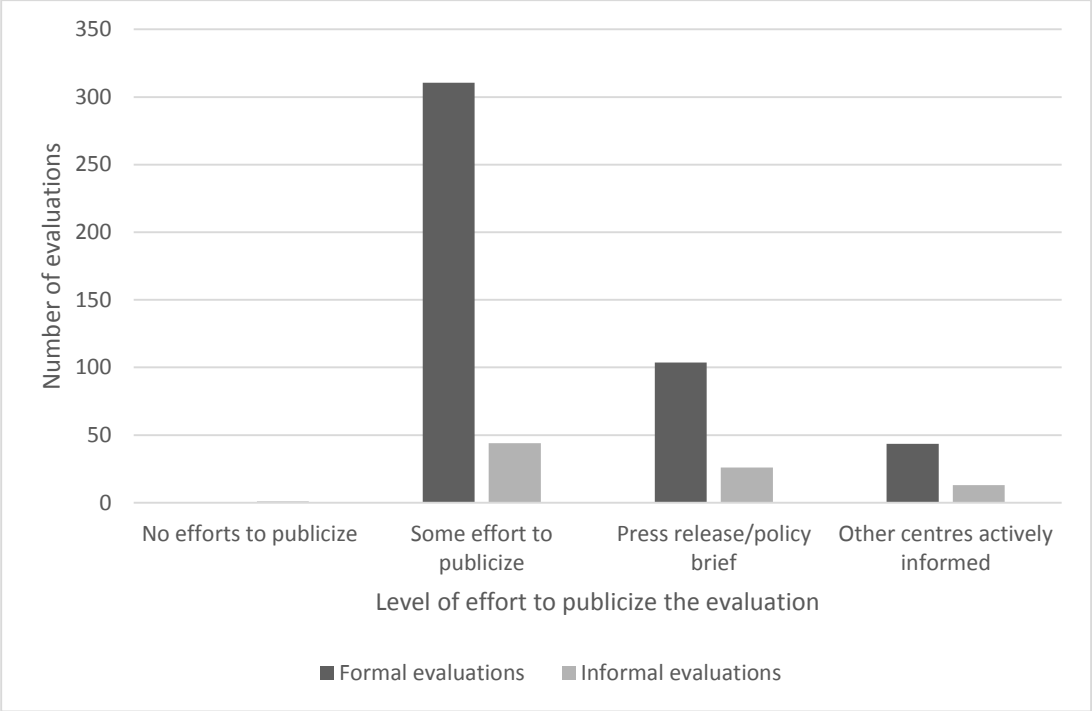
Figure 7.31: Formal and informal evaluations referencing each other (N = 542)[76]



Finally, as Chapters 5 and 6 have discussed, explicit efforts to publicize policy evaluations may have a bearing on the extent to which evaluative insights travel across governance centres. Figure 7.32 therefore presents the extent to which the authors of the evaluations made an effort to publicize their findings, for example by posting their evaluation on a website, issuing a press release, or informing other governance centres even more actively, such as through a conference or a workshop. Recall that this information was extracted both from the evaluation, but also from the source website or similar, where available (see Chapter 4).

[76] Extrapolation for formal evaluations.

Figure 7.32: Efforts to publicise formal and informal evaluations (N = 542)[77]



The first things to note in Figure 7.32 is that as Chapter 5 has already discussed, there are no formal evaluations that were not publicized at all, and only one informal evaluation where this was the case (the size of the chart is too small to represent that one evaluation – see Chapter 6 for a finer resolution). Proportionately speaking, 67.86% of the formally funded evaluators made some efforts to publicize (e.g., putting the evaluation onto a website), with 52.38% for informal evaluators. Only about a fifth (22.62%) of the formal evaluations were linked to a press release or policy brief (with 30.95% for informal evaluations), and only 9.52% of the formal evaluations were used to actively inform other governance centres (the analogous number being 15.48% for informal evaluations). Overall, Figure 7.32 thus reveals a slightly higher relative propensity to publicize findings among informal evaluations than among formal ones.

[77]Extrapolation for formal evaluations.

7.5 Conclusion

Comparing formal and informal evaluations reveals significant differences on some, but not all, of the dimensions of self-organization, context, and interacting governance centres. On the self-organization dimension, the first key point to make is that formal actors funded the vast majority of climate policy evaluations in the overall database (see also Chapter 4). In other words, the main impetus to fund evaluations comes from state actors or actors which draw significant funds from the state and are thus *not* self-organized according to polycentric governance theory. However, legal requirements are evidently not the main driver of the bulk of evaluations that formal actors funded – rather, the evaluations most commonly cited a desire to learn and improve climate policy. Furthermore, when focusing on the nature of the evaluation, it is also clear that formal actors are proportionately more inclined and/or able to financially support continuous evaluation exercises or cycles, whereas informal actors fund more ad-hoc evaluations.

Second, on the contextual dimension, informal evaluations score on average higher than formal evaluations on the context index containing eight individual variables, which shows that informal evaluations contribute a somewhat greater amount of evaluative knowledge on contextual factors into the polycentric governance system. This general trend also extends to other variables, such as the number of criteria, where again informal evaluations used a greater number. The only exception to this trend concerns the number of evaluation methods, where formal evaluations use on average more methods than informal evaluations.

Third, the category of interacting governance centres is probably least clear when comparing across formal and informal evaluations. By and large, climate policy evaluation in the EU focuses on single governance centres. For example, and linking with the section on self-organization, EU level funders mainly fund EU evaluators to evaluate EU level climate policy, with a very similar pattern emerging for evaluation funders in Germany and in the UK. Furthermore, while formal and informal evaluations are (proportionately) broadly similar with a view to the stated evaluation purpose, the target audience, the context contained in recommendations, the number and characteristics of executive summaries, linguistic access, filling gaps and publicity efforts, there are some notable differences: informal evaluations pay

more attention to other evaluations of the same centre (i.e. they tap into more sources of available evaluative knowledge); informal evaluations also contain more recommendations and use more data from formal sources than vice versa. By the same token, formal evaluations contain greater numbers of (quantitative) comparability metrics. Table 7.2 summarizes the key takeaway points.

Table 7.2: Formal and informal evaluation: key similarities and differences

<table>
<tr><th></th><th>Formal evaluation</th><th>Informal evaluation</th></tr>
<tr><td>Self-organization</td><td>• Vast majority (85%) of evaluation are not self-organized
• Legal requirements not the main driver</td><td>• 15% of evaluations self-organized
• Significant growth in numbers since 2010</td></tr>
<tr><td>Context</td><td>• Greater number of methods, quantification and continuous evaluation</td><td>• Greater attention to various contextual variables
• Use more social criteria such as fairness and legitimacy</td></tr>
<tr><td rowspan="2">Interaction</td><td colspan="2">• Strong congruence between location of funders, evaluators, and the climate policy under evaluation
• Learning as the headline evaluation purpose; followed by accountability
• Limited attention to insights from other governance centres</td></tr>
<tr><td>• More quantitative comparability metrics</td><td>• Pays more attention to findings from evaluations of the same centre
• More recommendations</td></tr>
</table>

Taking together these results show that while the potential for and actual interactions between governance centres by the means of evaluation is relatively weak for both formal and informal evaluation, the two types of evaluation have different characteristics on various dimensions. In other words, formal and informal evaluation each exhibit some characteristics that the other does not possess to the same degree; this indicates that climate policy evaluation has generally become richer and broader given the involvement of different types of evaluation funders as defined in this thesis. It also highlights that formal and informal evaluation are, at

least in some areas, potentially complementary. The following Chapter 8 provides a theoretical analyses of the findings from Chapters 5-7.

Chapter 8 Evaluation in Polycentric Governance: A Theoretical Analysis

8.1 Introduction

The core aim of this thesis is to understand the potential and actual role of policy evaluation in facilitating climate governance, by considering the former through the lens of polycentric governance theory (and vice versa). In so doing, it looks at policy evaluation in a new light: moving on from conceptualizing evaluation as the last stage in a nationally framed policy cycle, and starting from the premise that evaluation may weave much more profoundly into the fabric of governance. In many ways, the theoretical connection between polycentric governance literatures and wide-ranging policy evaluation literatures are debates on the role of different actors in climate policy evaluation (*formal* state-based actors and *informal* society-based actors, i.e. the *who?* question in Chapter 2). This link emerges because of the keen interest from polycentric governance scholars in the origins, maintenance, and structure of governance arrangements (i.e. does a polycentric governance system emerge from the top down or from the bottom up?) and the equally keen interest by evaluation scholars, who have recognized that it appears to make a difference what types of actors conduct policy evaluation in terms of results and use (see Chapters 2 and 3). This chapter uses the empirical findings from this thesis in order to discuss the core theoretical questions and debates related to polycentric governance and evaluation in detail. In other words, the chapter aims to relate the new empirical evidence in this thesis to the theoretical postulates of Chapter 2. As before, it considers the role of self-organization, context, and interaction in separate sections. After these three sections follows a fourth that looks across the three foundational ideas in order to analyse links and cross-overs. The final section concludes and bridges to the final chapter.

8.2 Self-organization

The first, foundational idea of polycentric governance concerns the extent to which governance actors are willing and able to self-organize and take action beyond what states and state-linked actors do. Recall that in line with arguments from polycentric governance scholars, this thesis distinguished between self-organization and its opposite, namely state-based organization, on the basis of evaluation funding (see Chapters 4; 5-7). This thesis started from a theoretical expectation that those who pay for an evaluation also have significant control over what is being evaluated and, to a certain extent, how and to what end (see also Pleger & Sager, 2016).

Empirically, self-organization may be considered in relation to funding and the contribution of individual governance centres. First, considering the entire database (see Chapters 4 and 7), the findings of this thesis demonstrate that the number of informal evaluations is comparatively small, as formal actors funded the vast majority of evaluations. In other words, this thesis demonstrates limits to self-organization in climate policy evaluation. Second, formal and informal actors at the EU level and in Germany funded policy evaluation in about equal numbers, with considerably less evaluation activity in the UK (see Chapters 4 and 7). This suggests that not all governance centres prove equally active in terms of funding climate policy evaluation. However, the proportion of informal evaluations (compared to formal evaluations) remains relatively similar across the centres; that is, a smaller overall number of evaluations in the UK does not affect the proportion between formal and informal evaluations there. This discovery suggests that there may be similar, underlying drivers for both formal and informal evaluation across different governance centres in the EU. Coupled with the fact that informal evaluations tend to emerge later than formal evaluations, this finding indicates that informal evaluations are more reactive, and less spontaneous and independent, than polycentric governance theorists have traditionally expected. Especially Elinor Ostrom (2005, p. 280) suggests that organizations may be able to combine their resources and conduct monitoring and evaluation, but this thesis only partially confirms this assumption (see Chapter 2). Not all governance centres and actors in the EU's polycentric climate governance exhibit equal capacities, propensity, or willingness to self-organize their policy evaluation.

The high geographical congruence between the evaluation funders, the evaluators and the location of the policy under evaluation furthermore reveals relatively separate spheres of climate policy evaluation in the EU, which coincide with the boundaries of the governance centres. In other words, evaluation tends to happen within individual governance centres and evaluations rarely look beyond their own governance centre for lessons from elsewhere. Significantly, this pattern holds across both formal and informal evaluation, producing considerable fragmentation in evaluative knowledge of climate policy effects (see below for a fuller discussion on related aspects of [the lack of] interaction). Self-organization in evaluation contributes somewhat to emphasizing certain climate policy sub-types in terms of the number of evaluations (notably emissions trading), but the overall sectoral focus of formal and informal evaluations remains similar. Both formal and informal evaluations mainly concentrate on renewables policy, cross-sectoral elements, emissions trading, transport and energy efficiency, while other climate policy sub-types such as biofuels, agriculture, or waste receive little attention. Self-organized evaluation thus does not cover key substantial and geographical gaps left by formal evaluations (and vice versa).

Polycentric governance systems are, by definition, assumed to be dynamic and ever-changing (Morrison, 2017). The experimentation at various governance levels that Chapter 2 discussed suggests high levels of dynamism and change in the search for adequate governance responses to climate change, and corresponding and equally dynamic evaluation efforts. This dynamism arguably calls for continuous evaluation that captures change over time and assess its effects (see Morrison, 2017). In this regard, self-organized evaluations make a weak contribution; informal actors funded proportionately fewer continuous evaluations than ad-hoc evaluations, while formal funders funded ad-hoc and continuous evaluations in nearly equal numbers (see Chapter 7). Thus, in addition to limits to self-organization in evaluation (see above), most informal evaluations do not exhibit and capture the dynamism that is assumed within polycentric theory. Some of the (rather optimistic) assumptions about self-organization may thus have to be tempered in the area of climate policy evaluation.

What do these findings mean for the aforementioned debate on the role of the state in polycentric governance? Recall that Chapter 4 explained that the formal (state) category contains evaluations funded by courts and scrutiny bodies,

parliaments, governmental organizations, banks, or agencies, independent advisory committees, research councils, and the (executive) government. This was done because, as Chapter 2 explained in greater detail, scholars working in the polycentric tradition tend to conceive of the state in a relatively broad way (see Mansbridge, 2014). The evidence from actor types among evaluation funders suggests that states play a key, but not an exclusive or exhaustive, role in polycentric climate governance through evaluation in the EU because they fund the lion's share of climate policy evaluation.

In addition to directly funding evaluation, states also play a role in putting in place rules for evaluation. Legal requirements to evaluate (or evaluation clauses) may be one way for formal actors to provide aspects of overarching rules in polycentric systems to ensure a minimum level of information provision vis-à-vis evaluation. Chapter 2 discussed earlier evidence of growth in the number of evaluation clauses and corresponding evaluation activities in Europe, but highlighted the need to study these dynamics in the climate policy sector. The new data in Chapter 7 demonstrate that states play only a partial role in this regard. While (unsurprisingly) none of the informal evaluations were triggered by legal requirements, it is especially notable that most formal evaluations were also *not* stimulated by legal requirements either; other factors, notably a desire for learning, proved more influential in stimulating formal actors to fund evaluations (see Chapter 7). In other words, in some respects, the role of the state is not as heavy-handed as it could be across parts of the formal evaluations because formal actors funded many evaluations outside some legal requirement to evaluate. This finding is somewhat at odds with the (more general) evaluation literature, where the presence of evaluation clauses has typically been understood as a sign of rising institutionalization of evaluation, and generally a growth area (Jacob et al., 2015; Mastenbroek et al., 2016). This thesis shows that evaluation may not (yet) be as institutionalized in the climate sector as in other sectors (such as structural funding policy in the EU, see Chapters 2 and 3).

Given that what counts as 'self-organization' depends so intimately on the definition of *formal* and *informal* evaluation, the point that both categories are much more internally differentiated than discussed in the relevant policy evaluation literatures is a novel finding, which has relevant theoretical implications for

polycentric governance. On the one hand, self-organized evaluations indeed contribute new insights on climate policy into the polycentric governance system; on the other hand, Chapter 6 has revealed that self-organization is limited to a relatively small number of organizational types, namely environmental organizations, industry, and foundations. The fact that public interest organizations, such as consumer protection groups or trade unions, funded virtually no climate policy evaluations indicates that the self-organization via evaluation observed here emerges from a small number of tightly circumscribed interest groups, such as the WWF or the federations of the renewable and solar energy industries in Germany. There are significant swathes of society or groups, such as human rights or public health organizations, that do not fund (and thus self-organize) climate policy evaluation at the EU level, in Germany, or in the UK. Those who do self-organize only engage in a limited set of evaluations (see Chapter 7). Particularly the more resource-intensive activities, such as using multiple evaluation methods or generating more elaborate quantifications of policy effects remain very rare in self-organized evaluations (see below). The same argument also holds for more continuous climate policy evaluations that are part of broader evaluation cycles which are rarely self-organized. This implies that, again, the pooling of resources among informal actors that Elinor Ostrom (2005, p. 280) proposed only happens to a limited extent, and mostly through certain organizations, such as the national or European-level offices of environmental organizations. These organisations produce mainly ad-hoc and thus less resource-intensive evaluations.

8.3 Context

The second foundational idea in polycentric governance theory is that context matters immensely so that policy solutions are unlikely to work everywhere once they meet widely differing contexts—hence repeated reminders to avoid policy ‘panaceas’ (E. Ostrom et al., 2007). The theoretical implication of this idea for policy evaluation is that any evaluation of policy effects should take contextual factors into account, lest it misses key mechanisms that lead to certain policy outcomes. Chapter 2 explained that evaluation scholars have proposed to account for context in two ways, namely directly assessing contextual factors such as time, the political

environment, or external shocks, and/or considering context through the use of many different methodologies and criteria. Paying attention to context is especially relevant for interactions between governance centres (see Section 8.4 in this chapter), which depends on contextual information if lessons are to travel from one governance centre to another without reverting to policy panaceas in polycentric governance systems.

Overall, the empirical findings from this thesis suggest that at best both formal and informal evaluations exhibit a cursory treatment of context. Both in terms of assessing contextual factors directly and looking at the number of methods or criteria that evaluations use, the average attention to context remains low across in climate policy evaluation. However, some important differences between formal and informal evaluations emerge in this thesis. Looking across the context index (see Chapters 5-7) reveals that informal evaluations pay, on average, more attention to contextual factors than formal evaluations. This suggests that informal evaluations not only exist (i.e. that informal actors have the capacity and willingness to evaluate, see Section 8.2 above), but that they are also *qualitatively* different on the contextual dimension. Informal evaluations contribute an additional contextual focus to the available evaluative knowledge on climate policy effects vis-à-vis policy evaluation; this knowledge is, on average, less available from formal evaluations. Thus, it is advantageous to have a significant presence of informal evaluations in a polycentric climate governance system, but their numbers are currently limited.

Considering other ways of gauging attention to context in climate policy evaluations – such as the methodological approaches and assessing policy targets – shows that formal evaluations use a greater number of methods. Compounded with the fact that formal evaluations also contain a greater number of quantifications on 'comparability metrics' (see Chapter 2), this indicates that formal actors may have higher levels of resources and/or willingness to fund evaluations that tend to be especially costly, such as using multiple methods or extensive quantification. This speaks directly to one of Elinor Ostrom's core claims that non-state actors may also have the capacity to conduct their own evaluations (E. Ostrom, 2005, p. 280). This part of the evidence suggest that it is indeed possible to empirically detect informal climate change policy evaluation at the EU level, in Germany and in the UK, but informal actors appear to have limits in the extent to which they are able and willing

to engage in more resource-intensive types of evaluation. Anything that requires significant resources – such as using a greater number of evaluation methods, or adding more quantifications to the evaluations – are areas where on average formal, state-based or linked actors such as the European Commission or national ministries appear to have a distinct advantage. This finding again suggests a substantial – though by no means exclusive or exhaustive – role for states or network-managing organizations like the European Commission in polycentric governance (see Chapter 2).

While informal actors may experience resource limitations, the formal actors appear to have political limitations that inhibit them to a certain extent from engaging in more context-sensitive climate policy evaluation. These limitations appear across several categories. As Chapter 7 explains, informal climate policy evaluations employ a greater number of criteria than formal ones, and are also significantly more reflexive, meaning that they exhibit a greater propensity towards critically questioning extant policy targets. In line with Per Mickwitz's suggestions, informal evaluations are thus more likely to unearth a wider range of climate policy effects in context than formal evaluations (Mickwitz, 2003). Retrospectively questioning policy targets (i.e. reflexivity) may come with considerable political risks, given that doing so can highlight the shortcomings of earlier policy decisions at the expense of the powers that be. From a polycentric theoretical perspective, doing so is absolutely vital if insights are to travel from one context to another, and especially if governance centres aim to draw on evaluations in order to set new policy targets (and if one does not want to suffer from a success bias). In fact, Chapter 2 discusses how some authors assume that "reflexivity is a systemic feature" (Aligica, 2014, p. 66) in polycentric governance. The empirical evidence in this thesis shows that the majority of evaluations fall well short against this expectation and that on account of evaluation characteristics, polycentric climate governance in the EU may not be as reflexive as polycentric governance scholars like to assume.

But this thesis also demonstrates that what counts as (relevant) context may at least in part depend on the characteristics of the (climate sub-) policy type. For example, a long-standing policy that has already been subject to multiple revisions may cry out for more historical, time-based contextualization. By contrast, a recently created and implemented policy on, say, the deployment of renewable energy

technology, may have less of a historical dimension, but rather link closely with more contemporary developments in political structures and institutions, as well as aspects of (physical) geography if the renewable resource is not available in the same way in all places. The type of policy instrument may also have a bearing on what contextual factors prove particularly relevant – for example, the strength of enforcement institutions may matter in the case of a policy that regulates carbon emissions, whereas in the case of a distributive policy such as subsidies for renewables deployment, the availability of investment capital may be a relevant factor in different actors' ability to benefit from the subsidy (see Lowi, 1972 for a typology of basic policy types). This means that assessing context co-evolves with the characteristics of the policy in question.

Given that there is practically an infinite number of ways in which one could assess context, contextual evaluation requires a (hopefully conscious) choice of factors that an evaluation assesses. But because any evaluation is necessarily value-based (Vedung, 1997) and therefore a political activity (Weiss, 1993), contextual analysis (and the choice of potentially relevant contextual factors) may also reflect the interests of certain actors, and the policy values that they seek to pursue. For example, different actors may actively contest what factors, elements, or aspects are central to a policy and which ones are considered peripheral. This, in turn, means that the information that may travel between governance centres is, to a considerable extent, infused with the values that are inevitably part of any evaluation. Policy evaluation in a polycentric system is thus also an exercise in assessing what policy values particular governance centres consider paramount in their climate policy decisions – and these values may then be expressed in terms of policy targets and the assessment of contextual effects. This thinking links back to Frank Fischer's (2006) idea of 'societal vindication,' or evaluating against the values that a society deems particularly relevant (see Chapter 2). Especially given political sensitivities in the EU Member States, network managers such as the European Commission may thus be better placed to conduct such evaluations.

In summary, the empirical analysis has shown that both formal and informal evaluations make unique and relevant contributions to climate policy evaluation in the EU. The evaluations analysed here contain many of the elements that Rog (2012) identified in her framework for contextualising evaluations (see Chapter 2).

However, this thesis also shows that there are additional factors that may prove relevant – such as the tailoring of evaluation methods or reflexivity against original policy targets. In other words, both formal and informal evaluations make relevant contributions, which are to a certain extent complementary in facilitating polycentric climate governance.

8.4 Interaction

The third foundational idea in polycentric governance theory concerns the interaction between governance centres (see Chapter 2). The potential for interaction via evaluation is in many ways a first and significant (but not yet sufficient) step in facilitating polycentric climate governance. Polycentric governance scholars have long argued that in the absence of any interactions between governance centres, it would be futile to speak of a governance 'system' (see Chapter 2). Chapters 4-7 have assessed the potential for interaction vis-à-vis climate policy evaluation in two broad ways: first, they considered to what extent certain evaluation characteristics may provide a base for interaction; second, interactions have been assessed with a view to the extent that they manifest directly through evaluations, for example, the use of data across different evaluations or the extent to which evaluations themselves engage with the experiences from other governance centres.

The fact that 'learning and improvement' is the most widely-cited reason for conducting evaluations – both in their formal and informal variants (see Chapter 7) – shows that most evaluators at least *aim* to stimulate learning and lesson drawing. Furthermore, the fact that most evaluations analysed here are geared towards policy-makers and politicians as their core audience (see Chapter 7) indicates an interest in more detailed learning, rather than just symbolic action. However, the empirical findings suggest that interaction between governance centres by the means of and directly through evaluation remains limited and typically does not extend much beyond relatively standard ways, such as executive summaries.

However, the analysis revealed important differences between formal and informal evaluations. Formal evaluations contained significantly more quantitative data and a greater number of comparability metrics (see Chapter 7). These

differences have important theoretical implications. Polycentric governance scholars have long debated the ability of self-organizing actors to muster the necessary and considerable resources to conduct scientific studies or, given the focus of this thesis, policy evaluations (Mansbridge, 2014; E. Ostrom, 2005). Recall that estimates from Germany have identified a cost of about one hundred thousand Euros per evaluation in the structural funds sector (Löwenbein, 2008). While the evidence in this thesis suggests that informal actors have indeed conducted evaluations in the climate sector, limits appear within these evaluations, as, for example, using various methods or quantifying policy effects more extensively are resource-intensive activities. In this respect, formal evaluations engage significantly more in these activities, perhaps because state-linked actors are typically backed by greater levels of resources.

But keeping in mind that Elinor Ostrom counselled extensively against policy panaceas and monocentricity (E. Ostrom et al., 2007), it is also relevant to note that formal climate policy evaluations contextualize less than informal (see above). Thus, even though they contain a greater number of quantitative comparability metrics and thus a means to carry some more general lessons, lower levels of contextual information may hamper the ability of potential evaluation users to determine the extent to which the policy context in other governance centres may be similar (in a familial way) with their own, and thus gauge to what extent the lessons may apply (see Chapter 2). This is in line with the thinking of both polycentric and evaluation scholars that contextual factors often have a strong bearing on policy success (Chapter 2).

These aspects apply specifically to policy recommendations, which are a fairly direct way of information exchange. Chapter 7 revealed that less than half of all evaluations in the database contained 'extensive recommendations,' which means that the actual contribution of both formal and informal evaluations to polycentric climate governance is limited with regard to providing recommendations. However, informal evaluations contained on average a greater number of recommendations, indicating that they contribute disproportionately more than formal evaluations. But neither formal nor informal evaluations evidently expended much effort to contextualize their recommendations – perhaps in order to underline their assumed generalizability. This means that potential evaluation users would have to go through

entire evaluations rather than just the recommendations in order to pick up on key contextual elements.

However, other aspects of the evaluations tell a somewhat different story: most evaluations contain executive summaries that distilled key findings – a characteristic that evaluation scholars had already identified as a key prerequisite of usability (Mastenbroek et al., 2016), and many of the executive summaries also have an internal structure in order to highlight the most relevant points (such as bullet points or figures). By contrast, the evaluations' ability to travel across language barriers in Europe is limited. Very few evaluations contain translations of their abstracts, let alone the whole evaluation (see Chapter 7), a phenomenon that stretches equally across formal and informal evaluations. The ability of climate policy evaluations to cross language barriers and thus governance centres therefore remains restricted and may indeed be one of the main obstacles to more interaction. This state of affairs excludes a range of actors from evaluation use and thus limits the accessibility of evaluative knowledge. This contrasts with Vincent Ostrom, who considered openness in polycentric governance especially paramount (see Chapter 2). However, given that many evaluations were written in English, it is of course also possible that policy elites across Europe and beyond will be able to engage with the evaluations even though they are not available in their native language (but his will unlikely hold for all members of the public). Finally, limited efforts to publicize evaluations to other governance centres is another indicator pointing in the same direction of limited interaction between governance centres through evaluation.

On the second point, namely interactions *through* evaluation, the fact that formal evaluations appear to engage less with other evaluative insights from the same governance centre highlights that self-organization and interactions between governance centres are by no means independent: informal evaluations contribute disproportionately more to engaging with other governance centres through their own analysis. Thus, informal evaluations proved somewhat more outward looking than formal evaluations. But at the same time, both formal and informal evaluations are relatively weak in identifying and addressing the gaps left by the other, although there is a very clear difference between them in the sense that informal evaluations are much more likely to use data and insights from formal evaluation than vice versa. There are also clear limitations among both types of evaluation when it comes to

engaging with the experiences of other governance centres (see Chapter 7). In other words, the climate policy evaluations considered in this thesis appear limited because evaluation mainly happens within individual governance centres (in jurisdictional terms, as well as to a certain extent along the formal-informal distinction). From this point of view, one could only speak of a very limited role for climate policy evaluations in linking insights from different governance centres and therefore enabling the flow of knowledge and exchange through their own analysis.

In summary, this analysis reconfirms that the concept of interaction between governance centres is by no means monolithic, as there are many potential ways in which climate policy evaluations may facilitate interaction (see Chapters 2 and 4). Both formal and informal evaluations make distinct and on the whole complementary contributions to interaction, but fall short of realizing their full theoretical potential. While formal evaluations appear to benefit from a higher level of resources that allow for greater methodological plurality and/or quantification, informal evaluations have strengths in other areas and can, for example, have a more integrative function because they engage more with other climate policy evaluations across governance centres.

However, the collective action problem on knowledge production keeps surfacing in this thesis; particularly activities that would primarily benefit other governance centres, such as contextualizing findings, providing translations or publicizing findings, feature weakly across the evaluations that this thesis analyses. While, for example, providing evaluations in different languages comes at a cost for the evaluation producers, the users are typically elsewhere and may, to a certain extent, 'free ride' on the insights of others – a phenomenon that Elinor Ostrom had already identified as one of the potential shortcomings of polycentric climate governance (see E. Ostrom, 2010c). Even the fairly basic task of pooling all (or at least most of) the evaluative information and looking across it (in effect the task that this thesis begins by assembling a novel evaluation database) was previously neglected, meaning that there is a wealth of climate policy evaluations that had not been brought together systematically—although there are some limited databases (Schoenefeld & Jordan, 2017), including those used as a data source for this thesis (see Chapter 4). To date, formal and informal evaluation only contribute partially to

addressing the collective action problem of creating and synthesizing the necessary evaluative knowledge on climate policy effects.

8.5 Looking across the three foundational ideas

The chapters of this thesis are structured around the three foundational ideas of polycentric governance – namely self-organization, context, and interaction. And yet, especially the sections above suggest that they are intimately linked in practice. This section aims to address these linkages more specifically, drawing on the evidence discussed above.

First, there is the link between self-organization and context. The empirical evidence indicates that the self-organizing, informal evaluations tend to contextualize their findings more than formal evaluations. While an observation of a correlation cannot provide a fully satisfactory *causal* explanation of a possible connection, several additional aspects point towards a possible causal relationship. Chapter 7 has revealed that the difference in contextualization on the index score is mainly driven by greater attention to other sectors, unintended outcomes, and the political and institutional environment. These are, in many ways, areas of evaluation that could ultimately prove highly uncomfortable for formal actors, because they reveal unanticipated shortcomings of policy (i.e. potentially highlighting the limits of policy knowledge and thus decision-making), and because they address the nature of formal, state actors and governance structures themselves. Addressing these issues would require deep self-reflection and potential change, all of which are possibly risky activities that political actors may be keen to avoid. Further support for this argument emerges from the fact that informal evaluations also appear more likely to critically question extant policy goals; that is, they are, as Chapter 2 has highlighted, more reflexive. Because questioning earlier policy targets may upset existing political arrangements and actors, doing so may not feature strongly in formal evaluations.

Second, there is a connection between self-organization and interaction between governance centres. In general, this link emerges in some areas, but is not quite as strong as in the other cases (see above and below). While informal

evaluations exhibit a somewhat greater propensity to integrate knowledge from numerous climate policy evaluations (see above), on many of the other variables, there are no significant differences between formal and informal evaluations. Aside from relatively standard activities, such as providing executive summaries in evaluations, both formal and informal evaluations evidently struggle or are unwilling to engage in the more resource-intensive activities to foster interactions in an active and direct way such as providing translations or engaging in active publicity of their results. From a theoretical standpoint, this links back to one of the most fundamental, underlying premises of polycentric governance, namely collective action dilemmas and the provision of public goods. Polycentric governance assumes interactions between governance centres, but the willingness and ability to foster these via evaluation of climate policy in the EU is evidently limited, both among state and societal actors. Given that much of the requisite information already exists in the form of evaluations at the EU level, in Germany and in the UK, the need for information clearing houses, such as databases, becomes ever more evident (see Schoenefeld & Jordan, 2017). Relatedly, the former President of the American Evaluation Association Debra Rog wrote back in 2012 that

> We are increasingly cognizant that the work we do in any single evaluation should have cumulative force. Single studies are rarely definitive, but often fit within a broader literature (Rog, 2012, p. 37).

But it is of course an open question who 'should' or 'could' conduct these activities – an issue that links back with the idea of a potential network manager (Jordan & Schout, 2006; Jordan et al., 2018). One international actor that has already taken significant steps towards such network management is the Global Environment Facility by funding the Climate-Eval (now Environment-Eval) network.[78] But as Schoenefeld and Jordan (2017) highlight, this network is only a single attempt to bring together knowledge from different evaluations. Compiling evaluations in a database is probably a necessary first step, but does not, in itself, let the information flow (the evaluations need to be used and potentially analysed for this to happen).

[78] https://www.climate-eval.org/

It is also relevant to ask at what scale it makes most sense to do so – should European-level or national actors (such as the European Commission, the European Environment Agency or the German Federal Government) bring together information in a database and thus act as network managers, or is this best done at the global level (see Mickwitz, 2006, p. 71)? The findings from the previous chapter suggest some 'nesting' of climate policy evaluation, in that it detected evaluation activities at both the EU level, as well as the national (Germany and the UK) level. In sum, interaction is fundamental for polycentric governance, but self-organization – as operationalized in this thesis – is making at best a very limited contribution to fostering it.

Third, there is the link between context and interacting governance centres. As Chapter 2 has explained, this link affects the ability of lessons to travel from one governance centre to another in meaningful ways. Recall that Rog (2012, p. 37) highlighted that "[a]ttention to the contextual elements […] may help to make the [evaluation] findings more generalizable." Attempts to apply far-reaching policy prescriptions without attention to context could amount to the 'panacea thinking' against which Elinor Ostrom and colleagues (2007) counselled extensively. The evidence in Chapter 7 suggests that while the general level of contextualization tends to be low in all evaluations, this is especially true when highlighting contextual aspects in policy recommendations, which is one of the more salient places where actors from other governance centres may look for lessons and inspiration. Overall, low levels of contextualization in many areas show that from a polycentric perspective, there are significant ways in which climate policy evaluation practice – and especially its outputs – could be improved.

8.6 Conclusion

Two key conclusions have emerged. There is indeed an emerging *theoretical* rationale for a role for *ex-post* policy evaluation in facilitating polycentric governance systems, especially in the case of climate governance. Evaluation may contribute to the flow of information between governance centres if its characteristics

support this function. This flow of information is, in turn, a key way in which polycentricity in climate governance may turn into polycentrism (see Chapter 1).

Second, the empirical explorations that followed have shown that some, but not all, of the expectations derived from polycentric governance theory have materialized in climate policy evaluation output produced in the EU. While some of the evaluations were self-organized and others were not, a range of factors appear to support a linking function of evaluation across governance centres, but there are also key gaps on other dimensions. Policy evaluation would certainly be poorer without the self-organized evaluations, but the inverse also applies: formal (i.e. non-self-organized) policy evaluations play an important role in generating and potentially diffusing knowledge in a polycentric climate governance system. Especially the question of collective action dilemmas – arguably the core point of departure of the polycentric approach (E. Ostrom, 1990) – has emerged as an important issue with regard to knowledge production and provision via evaluation. In other words, while self-organization may assist in resolving some of the collective action dilemma in the area of evaluation, it is certainly no panacea, and states will likely continue to play a key role in stimulating and providing evaluative insights. However, those writing on polycentric governance are less clear on the balance of effort between state and non-state actors. While Elinor Ostrom has typically advocated a mix between the two, others have interpreted her work as an anti-state message, and thus spurred considerable debate (see Mansbridge, 2014). Evaluation may enable the systemic functions of polycentric governance by allowing lesson-drawing across centres without doing so in a way that ignores context, politics, values, and a range of factors that may help or hinder such a role. The state is likely to play a significant role in this process.

This thesis uses three prominent governance centres, namely the EU level, Germany and the UK, as its empirical focus. The results show that evaluation output differs across each centre, both in numbers and in nature. In addition, there are notable differences between formal (i.e. state-funded) and informal (i.e. society-funded) evaluation within each centre. This variation is expected – polycentric governance theory assumes different (evaluation) practices, and indeed heterogeneity and experimentation in different governance centres. These differences may ultimately even become an opportunity for learning about different ways of

organizing and practicing policy evaluation (see Schoenefeld & Jordan, 2017 for a related argument). The next and final chapter discusses the implications of this thesis from a broader perspective and starts to look ahead, both conceptually and in terms of policy recommendations and future work.

Chapter 9 Conclusions and New Directions

9.1 Introduction

After the failure of the 2009 Copenhagen summit to extend the Kyoto-based mechanisms to address climate change, many scholars suggested alternatives in the polycentric governance tradition (E. Ostrom, 2010c; see also Cole, 2015). But it soon became clear that what Elinor and Vincent Ostrom and others were proposing was neither theoretically fully specified, nor sufficiently empirically validated (Jordan et al., 2015). In and of itself, this situation already constituted a significant research gap, but the fact that the 2015 Paris Agreement ushered in an even more polycentric way of governing climate change at the international level (Oberthür, 2016) has further exacerbated the need for more theoretical and empirical exploration. In response, the core aim of this thesis is to understand what can be learnt about the potential and the actual role of one factor, namely policy evaluation, in facilitating polycentric governance. This aim led to two core objectives:

Objective 1: To identify the key foundational ideas of polycentric governance theory and relate these to relevant debates on policy evaluation in order to understand the *potential* role of evaluation in facilitating climate governance.

Objective 2: To test these theoretical expectations in the case of the European Union in order to understand the *actual* role of evaluation in climate governance.

This chapter returns to the original research aim and to the objectives, discusses the findings with a view to debates in polycentric governance and policy evaluation literatures and then relates these insights to the broader research context. The chapter closes with policy recommendations and reflections on promising avenues for future research.

9.2 Reflections on the original aim and objectives

9.2.1 The research aim

The overall aim of this thesis is to understand precisely which factors enable polycentric governance systems to function, with a particular focus on the potential and actual role of policy evaluation. Scholars have often assumed that knowledge emerges and flows in polycentric governance systems, but the role of evaluation as an important source of knowledge in governance had not yet been fully explored.

What have we learned from this project? This thesis is the first, to the knowledge of the author, to identify a substantial theoretical role for policy evaluation in facilitating polycentric climate governance. In developing these insights, this thesis significantly clarifies, specifies, and adjusts polycentric governance theory in order to study policy evaluation and generate a basis for empirical testing. Doing so demonstrates that in theory, evaluation amounts to more than a simple checking device at the end of a stylized policy cycle. In order to test the theoretical expectations from polycentric governance theory, this thesis presents empirical evidence from a novel database of 618 climate policy evaluations from three governance centres in the EU, namely the EU level, Germany and the UK. The following sections return to the two specific objectives of this thesis, namely theoretical development and empirical testing, in greater detail.

9.2.2 Objective 1: Theory development: polycentrism and evaluation

In the original reading of the theory, this thesis uncovers that polycentric governance builds on three core foundational ideas with crucial relevance for policy evaluation: (1) that actors have the willingness and ability to self-govern, (2) that context matters in governance, and (3) that governance centres, while formally independent, interact in order to harness the benefits of polycentric governance. This thesis specifies the three foundational ideas for the first time in order to connect with and structure existing debates in policy evaluation literatures (which had also begun to discuss the role of different evaluation actors, how context matters in evaluation, as well as to a lesser extent, thoughts on interaction and lesson-drawing as a function

of evaluation and thus hold important insights for polycentric governance and vice versa). This thesis is the first to develop a theoretical rationale for how the polycentric governance perspective helpfully illuminates the relationships among various debates in policy evaluation literatures (e.g., how self-organization and attention to context hang together), and what gaps still exist in theoretical and empirical explorations.

What core theoretical lessons does this thesis derive about how evaluation facilitates governance from a polycentric perspective? First, rather than assuming that policy evaluation would be the exclusive domain of formal state actors, the polycentric perspective stresses the possibility that many different actors may fund and/or conduct evaluation, some linked to the state and some not. In other words, self-organized policy evaluation is an important theoretical focus, and thus something to explore in detail. Evaluation literatures had also already engaged with the idea of different actors in evaluation, particularly with a view to the level of independence of different evaluation actors. A distinction between formal, state actors and informal, societal actors had already entered policy evaluation literatures as early as the 1970s, and had been sporadically discussed in the intervening decades (see Chapter 2). This thesis re-kindles and systematizes that discussion given its importance from the polycentric perspective. Crucially, it highlights that the formal-informal distinction turns out to be much more complex than first meets the eye because evaluation includes funders and evaluators, which can, but do not necessarily have to be, the same organization or individual. In order to capture the original impetus for evaluation, this thesis distinguishes between formal and informal evaluation on the basis of funding (rather than the nature of the evaluator, or whether or not an evaluation responds to a legal requirement). This highlights a complex layering of interests within evaluation endeavours, and the need to further explore the impetus for evaluation and its effects.

The second foundational idea of the polycentric perspective emphasizes the importance of context in public policy (and thus in evaluation), but it also recognizes the existence of contextually-embedded lessons that may emerge from evaluation. This thesis reveals that both polycentric governance and policy evaluation literatures had already dealt with the vexed theoretical issue of context in governance, but in somewhat different ways. While the importance of context in determining policy

outcomes is in many ways one of the starting premises of polycentric governance (after all, if context did not matter, monocentric solutions would suffice), policy evaluation scholars had long debated whether public policies generate effects irrespective of their context on one extreme end of the spectrum, or whether policy-making is so deeply context-dependent that the only way to understand it is from a historical perspective (see Chapter 2). Across the two literatures, those working on the issue of context turned out to be rather united in their view that 'policy panaceas' are unlikely to work regardless of the context (see E. Ostrom et al., 2007). Theoretical discussions (see Chapter 2) have revealed that context is often idiosyncratic, making it difficult to define relevant contextual aspects *a priori*. But even though context is thus highly multi-dimensional, evaluation literatures had nevertheless identified various aspects that prove relevant for climate change. Therefore, this thesis unpacks a number of potentially relevant contextual factors for climate change such as the nature of political institutions, external shocks or geography, but combines these factors with attention to indicators or metrics that may carry more general, albeit context-embedded, lessons.

Third, the precise mechanisms for how governance centres interact and what role evaluation may play in that process had not been fully theoretically specified. But both the polycentric and the evaluation literatures have something to offer: while polycentric governance calls for interacting governance centres (but falls short on fully specifying the exact means to do so), policy evaluation provides a potential means but the respective literatures mainly focus on evaluating individual policies and by and large neglect larger governance issues. For evaluation, polycentric governance opens up the possibility that evaluation results may not only matter for individual policies, but that such insights could be relevant elsewhere. Such ideas have recently been floated in the evaluation community (Uitto, 2016). In linking polycentric governance and policy evaluation, the idea of lesson-drawing via evaluation becomes a distinct theoretical possibility from the polycentric perspective (see Chapter 2). This thesis specifies how evaluations (and the lessons therein) would have to look in order to facilitate interactions via lesson drawing. Important factors include contextual information, and lessons as well as potentially quantitative metrics that can carry more general insights, in addition items such as executive summaries and translations.

9.2.3 Objective 2: Testing the role of evaluation in climate governance

This thesis contains a new database of climate policy evaluations at the EU level, in Germany and in the UK (both national level). A systematic collection returned 618 climate policy evaluations (from 1997–2014), more than twice what Huitema et al. (2011) had found between the 1990s and 2007. After a ten year hiatus and with fast-moving debates on the role of policy evaluation in climate governance (Aldy, 2014; Aldy & Pizer, 2014; Aldy & Pizer, 2015; Aldy, Pizer, & Akimoto, 2017; Feldman & Wilt, 1996; Fransen & Cronin, 2013; Hildén et al., 2014; Hildén, 2014; Mela & Hildén, 2012; Schoenefeld & Jordan, 2017; Schoenefeld et al., 2018) the database is a crucial first step in anchoring these debates in more empirical evidence and also reflecting the fact that some of the most significant climate policies in the EU were put in place between 2007 and 2009 with the adoption of the 2020 Climate and Energy Package (Jordan et al., 2010). It should also be noted, however, that the EU is in many ways a unique case in the sense that it exhibits both ambitious climate change policy, as well as evaluation capacities. However, by choosing Germany and the UK (national level) as the two country cases, this thesis cannot account for greater variation within the EU (e.g., differing evaluation capacities and climate policy ambition within other states, such as the new member states), or indeed differences within Germany and the UK at the regional or local level (see Schoenefeld & Jordan, 2017; Schoenefeld et al., 2018). These are potential areas for future research (see below).

Looking across the entire database reveals that formal, state based actors funded the overwhelming majority of climate policy evaluations in the EU (458 in total). Informal (i.e. society-based) actors only financially supported a much smaller number of 84 evaluations (see Chapter 7).[79] There is thus only limited self-organization in climate policy evaluation. This finding only partially confirms the expectation from polycentric governance theory that actors are able and willing to self-organize their evaluation. Therefore, spontaneity in governance via evaluation is limited. But in line with the theoretical expectations from polycentric governance

[79] For the remainder of the evaluations, the funder could not be determined – see Chapter 4.

(see Chapter 2), the 84 self-organized evaluations allow meaningful empirical analyses using the newly developed coding scheme (below).

It should however also be noted that in addition to distinguishing formal and informal evaluations by the nature of the funder (for the full justification, see Chapters 1 and 2), this could have alternatively also been done by looking at the organizational nature of the evaluator. While this thesis assumed that funding an evaluation allows significant control over the evaluation outcome, it is also in principle possible that evaluators insert their own perspectives and evaluation priorities, especially if they can muster creative ways to deviate from the demands of their principals/funders (see Pleger & Sager, 2016). Considering the data in this thesis from the latter perspective would have generated vastly different results – for example if 'Research institutes', 'Commercial consultancies', 'Industry Groups', 'Civil Society Organizations', and 'Environmental Organizations' were classed as 'informal evaluators' (see also Huitema et al., 2011), then 82.8% of the evaluations counted in Figure 4.5 (i.e., the entire database) would have been considered 'informal.' In this case, Elinor Ostrom's arguments about the self-organizing capacity in evaluations would have received far greater support than is currently the case in this thesis. Conducting the subsequent analyses of this thesis with this alternative distinction between formal/informal and potentially comparing the results to the analyses flowing from the approach in this thesis could also generate new results regarding the principal-agent questions in evaluation.

A second, significant finding reveals a close entanglement between the locations of the evaluation funders, the evaluators, and the climate policy under evaluation. Evaluation tends to happen within individual governance centres, both in terms of the actors that fund and ultimately conduct the evaluation, as well as in terms of the data that they use and where they look (or not) for additional insights on the policy under evaluation. Looking beyond one's own governance centre by the means of evaluation is the exception rather than the norm.

Analysing formal and informal evaluations with the coding scheme (see Chapter 4) reveals some similarities, but also differences between formal and informal actors, thus pointing to (some) heterogeneity in evaluation, as polycentric governance scholars would expect. Formal and informal evaluations are similar in terms of their focus on climate policy sub-sectors (such as renewables policy or

emissions trading, while neglecting for example agricultural policy) but they differ on other dimensions such as evaluation criteria, purposes, and methods—even within the single field of climate policy. Recall, however, that the evaluation of adaptation efforts was not included in this study – doing so may have further increased the diversity of evaluation, given ongoing debates about evaluating adaptation (see e.g. Dupuis & Biesbroek, 2013). Furthermore, the exclusion of evaluations that were not sufficiently systematic, or evaluations that are not publicly available (see Chapter 4), could also lead to underestimations of evaluation diversity in this thesis. But turning back to the data analysed here, evaluations by informal actors tend to contextualize (somewhat) more, but those by formal actors engage in more resource-intensive evaluation activities, such as using many different evaluation methods or quantifying their findings. Limited contextualization of policy recommendations in both informal and especially formal evaluations indicates that there is still a good deal of 'panacea thinking' (E. Ostrom et al., 2007) in the climate policy evaluation community, reducing the contribution of evaluation in realizing the theoretical benefits of polycentric climate governance because actors in other governance centres may struggle to assess to which extent the recommendations apply to their own context.

The latter point relates to interaction between governance centres, where the empirical evidence suggests that the bulk of climate policy evaluations remain rather 'insular' because they tend to focus on their respective governance centres, and do not directly engage with evaluations of other governance centres. While most formal and informal evaluations engage in relatively standard activities such as providing executive summaries or some level of recommendations (which provides some basis for interaction), there is much less evidence of activities such as carefully contextualizing policy recommendations or providing translations of the summaries or entire evaluations into other languages. Informal evaluations exhibit a greater propensity to draw on and thus engage with the findings from other evaluations. True to the theoretical claim that more assessment of potential contextual factors is advantageous in a polycentric setting, the existence of both formal and informal evaluation facilitates polycentric climate governance. However, climate policy evaluations in the EU are only partially equipped to enable the flow of knowledge between governance centres via evaluation, even though informal evaluations make a slightly stronger contribution in this regard.

9.2.4 Summary

This thesis demonstrates that there is much to be gained from looking at evaluation's role in climate governance from the polycentric perspective, both theoretically and empirically. Conceptually, the polycentric perspective proves useful in perceiving and disentangling various perspectives on policy evaluation that have so far only received limited attention. It is certainly true that polycentric governance remains "a veritable work in progress" (Aligica & Sabetti, 2014a, p. 5), but this thesis uses an approach that not only furthers existing theorizing, but also opens up many additional venues for further work. In introducing the polycentric governance approach, Chapter 2 highlighted how the Ostroms and their collaborators stressed "heterogeneity, diversity, context, and situational logic as critical elements in the analysis of institutions, governance, and collective action" (Aligica, 2014, p. 5). The work discussed above has – to the extent possible – attempted to work in this tradition, and it demonstrates the diversity and other relevant aspects with regards to climate policy evaluation at the EU level, in Germany and in the UK and its role in facilitating climate governance. The empirical findings suggest that the current characteristics of climate policy evaluation equip it only partially to fulfil this role.

9.3 Contributions to knowledge

The theoretical and empirical sections of this thesis develop new theoretical insights on the role of evaluation in facilitating polycentric climate governance and test them in a particular case, namely climate policy in the EU. Some, but not all of the theoretical expectations could be empirically validated, as for example self-organization in evaluation remains limited. This section explores some of the broader contributions of these insights to the various original strands of literature first discussed in Chapter 2.

9.3.1 Contributions to the policy evaluation literatures

In the past, policy evaluation literatures have either discussed and developed evaluation methods and content (e.g., Mickwitz, 2003; Patton, 2008; Pawson &

Tilley, 1997), or have they focused on the institutions (e.g., the European Commission) and the actors who advocate, conduct, and use evaluation (see Mickwitz, 2013; Schoenefeld & Jordan, 2017). The polycentric governance perspective helps to recombine various existing strands of the policy evaluation literature into a more comprehensive theory-driven framework. This approach contributes a fresh perspective to the evaluation literatures because it accounts for why numerous evaluation characteristics are relevant not only in and of themselves (or for more circumscribed theoretical aspects in debates on evaluation), but rather for the functioning of polycentric governance as a greater whole. Polycentric governance theory is thus a useful tool for re-ordering and systematizing numerous debates in evaluation literatures and their corresponding empirical explorations.

This thesis is the first to research the need to distinguish between formal and informal evaluation on the basis of evaluation financing (rather than the nature of the actors that ultimately conduct the evaluation). In so doing, it demonstrates that there are important and underappreciated connections between the actors that fund and/or conduct evaluations and the evaluation methods and content. For instance, formally-funded evaluations quantify more and use a greater number of methods, while informally-funded evaluations tend to use more criteria and produce, on average, more recommendations. These links have consequences for the role of evaluation in polycentric governance systems given that not all actors appear to produce evaluations with the same characteristics.

These insights chime with an emerging debate that evaluation often involves multiple actors and thus principal-agent relationships; after some early empirical explorations (see Hayward et al., 2013), these relationships have only recently attracted more explicit theorizing (see Pleger & Sager, 2016). These dynamics also appear in the climate change sector. There are no unitary actors with unitary motivations behind most climate policy evaluations; at the very least, there are potentially complex principal-agent relationships underlying the production of climate policy evaluations. What constitutes 'formal' or 'informal' evaluation is furthermore closely bound up with a definition of the state and what constitutes 'state' and 'non-state.' For example, many civil society organizations receive public funds, or connect with state institutions in one way or another (see Greenwood, 2011). This thesis makes concrete, empirical suggestions on how to better distinguish

between formal and informal evaluation. This distinction is important because both types of evaluation are qualitatively different on dimensions that matter for the facilitation of polycentric governance. However, alternative ways of distinguishing between formal and informal evaluation (such as focusing on the nature of the evaluators) could unearth additional insights, for example to what extent evaluators may be able to strengthen or counteract the influence of their funders (e.g., are mainly society-based evaluators still relatively independent when using government funds for an evaluation?). Furthermore, it may have been somewhat more straightforward to identify the organizational category of the evaluating organization (as the evaluator is normally mentioned in an evaluation), as opposed to the funder (note that doing so may have reduced the substantial number of evaluations – 74 in total, or 11.97% of the database). Thus, more data would have been available for analysis.

One important area of difference between formal and informal evaluation is context. Three contextual variables, namely attention to other sectors, unintended side effects, and the political environment mainly drive the difference between formal and informal evaluations in considering the context of the policies they study. Informal evaluations contextualize more in areas that are especially political, and where governmental actors may not wish to shine a strong spotlight. These areas are especially political because unintended side effects may reveal the inability of public actors to anticipate or control a range of policy effects; they may also not appreciate learning about such effects in other sectors, or being exposed to argumentation on the political environment, which they inhabit and know well. Policy evaluation in climate governance thus differs from the earlier discussions on monitoring in common pool resource systems advanced by the Ostroms. Elinor Ostrom typically wrote of the local context, such as the nature of the resource being managed (it makes a difference whether one is managing a population of lobsters or a pasture), the specific techniques used for monitoring as a function of the resource and so forth, but she put less emphasis on political elements. This thesis shows that the latter become an important ingredient of lesson-drawing in polycentric governance, but formal actors are not very well inclined to provide them in relation to climate policy.

In sum, this thesis proposes to conceptualize and study evaluation as an integral part of governance processes – a practice of enlightenment on policy effects

that is, by its very definition and ultimate purpose, a value-driven and therefore political endeavour. While evaluation is about judgement, polycentric governance is about structure and ordering of governance, as well as the related processes of experimentation, competition, and innovation. This thesis demonstrates how the 'valuing' in evaluation becomes relevant for polycentric governance. Evaluation can answer questions like 'Why did they do what they did in their governance centre? What values was a climate policy based upon, what visions of politics and its institutions and to what extent do these experiences bear relevance and lessons for other governance centres?' Understanding evaluation in this way moves significantly beyond considering 'one evaluation at a time,' and investigates the cumulative effect of numerous evaluations (see Rist & Stame, 2011). This is a much-needed perspective in times of rising evaluation output, but little theorization and empirical exploration of its broader governance role.

9.3.2 Contributions to polycentric governance theory

Numerous scholars had made the point that polycentric governance theory—and especially climate change governance theory—would benefit from much more theoretical explication and empirical application. While Aligica and Sabetti (2014a, p. 5) wrote that "the Ostroms left behind a veritable work in progress", Elinor Ostrom (2014b, p. 84) herself stressed that "a complete inventory […] [of climate governance efforts] would be a good subject for a future research project." Furthermore, Jordan et al. (2015) emphasized that the emergence of increasingly polycentric governance arrangements in the area of climate change governance cries out for empirical exploration, because many of its underlying assumptions had only scarcely, if at all, been tested.

This thesis contributes knowledge on a growing, but so far largely neglected activity with high relevance to polycentric governance: policy evaluation. In so doing, this thesis disentangles and refines three foundational ideas of polycentric governance. These refinements demonstrate that, first, 'self-governance' takes on a new and different meaning in international climate governance arrangements where states, interest groups, industry associations, and other types of organizations dominate. The constitution of the 'self' becomes more difficult to apply. In other

words, who is the 'self' that is monitoring or evaluating? Is it enough to have non-governmental or civil society organizations conduct policy evaluation? In the German context, evaluators had already been debating this very question (Struhkamp, 2007). Because policy evaluation is typically conducted by organizations or institutions that involve many individuals and that have existing (political) interests and relationships with other institutions, the idea of self-organization can only be applied in a very general sense to evaluation. Whereas in the case of common pool resource governance systems the interests of the individuals monitoring fisheries or water management are relatively clearly defined and generally in line with the interests of the overall system (i.e. they aim to ensure that the appropriation of common resources remains within self-determined limits), in the case of climate policy evaluation, such an alignment of interests cannot readily be assumed. The involvement of different institutions and/or individuals in evaluation cannot guarantee a unitary interest among all those involved in financing, conducting, and/or using the evaluations (see Pleger & Sager, 2016). These principal-agent dynamics are therefore a key area for future work in the context of polycentric governance (see below).

The distinction between formal and informal actors based on evaluation funding (see Chapter 4), allows exploring the role of the state in climate policy evaluation and, by extension, in polycentric climate governance. Overall, the overwhelming majority of formal evaluations (compared to informal evaluations - see Chapter 7) reveals *de-facto* a strong role for the state in climate policy evaluation in the EU. States finance and/or conduct most of the climate policy evaluation in the EU. But this is by no means an exclusive role, as informal (i.e. non-state) actors are also involved. Their involvement emerges not only in terms of the number of evaluations that they produce, but also in *how* they evaluate. Especially the more 'political' elements that are often neglected in formal evaluations are more central in informal evaluations (see above).

Both scholars arguing for a role of the state in monitoring and evaluation activities (see Mansbridge, 2014) and those with a more sceptical orientation (E. Ostrom, 2005) have thus identified relevant actors in polycentric climate governance whose activities turn out, in the end, to be rather complementary than antagonistic in the case of climate policy in the EU (see Chapter 8). From a polycentric perspective,

greater pluralism in climate policy evaluation in terms of criteria, methods, and actors resulting from the involvement of both state and non-state actors ultimately means broader (but as this thesis demonstrates not necessarily comprehensive) evaluative coverage of the many potential effects of climate policies, and this is where non-state actors contribute significantly to the variety of evaluation.

Furthermore, Chapter 2 reveals that the means of interaction between governance centres had remained, by and large, under-specified in existing polycentric governance literatures. Scholars in the polycentric tradition often highlight the (normative) need for such interactions, or implicitly assume their existence, but had not yet done sufficient work to test their existence and nature empirically. Vincent Ostrom writes about individuals that may serve as bridges between governance centres and Elinor Ostrom floats the idea of newsletters (E. Ostrom, 2005, p. 280) to carry experiences from one centre to another. But these discussions remained too general and too unspecific to derive clear findings. This thesis shows that policy evaluation has considerable theoretical potential to contribute to such interactions, but that in light of the empirical evidence, this potential has at best only been partially developed in the EU.

The idea of interaction between governance centres probably comes closest to the Ostroms' thinking in its application of polycentric governance theory to climate change in this thesis. Yet, it has also not been easy or straightforward to adjust the concept. First, there is the vexed question of who or what qualifies as a 'governance centre' (see Chapter 2). This thesis has pragmatically defined three governance centres as the EU level, Germany (national level) and the UK (national level), especially because the national level is where the bulk of the literatures on policy evaluation concentrates, although there is some 'nesting' of evaluation activities (see Chapter 3). The thesis demonstrates that key concepts from polycentric governance theory can be usefully adapted to this operationalization, but this is by no means the only level where insights on the role of policy evaluation apply. What defines an interaction and where that interaction happens is, again, an area of definitional and empirical exploration, and what has been done in this thesis is one, but certainly not the only way to do so. Explorations across other types of governance centres and across other levels (such as the regional/local and the international level), or the links between different 'domains' of polycentric governance such as the 'transnational'

and the ‘nation state’ domains as a function of policy evaluation could be further explored (Jordan et al., 2018).

There are also important questions on the idea of a higher-level jurisdiction (see Chapter 2). This thesis shows that states play an important role in funding evaluations, and therefore providing a source of information (Mansbridge, 2014), but this is by no means an exclusive or exhaustive role. Actors at the EU level were by and large the only ones to venture beyond their own governance centre in terms of funding evaluators in other governance centres. Therefore, the EU level appears to fulfil some of the functions of a higher-level jurisdiction in climate policy because it looks across several governance centres, but it also leaves important gaps. Crucially, neither formal nor informal actors played strong roles in systematically bringing evaluative knowledge together in the form of databases or other knowledge-sharing platforms (Schoenefeld & Jordan, 2017). In other words, a significant task ascribed to higher-level jurisdictions by the theory – namely bringing evaluative insights from different governance centres together in useful ways – remains, by and large, unfulfilled in the EU, and therefore detracts from realising the putative benefits of polycentric governance. This significant shortcoming is notable for the EU, which has long sought to portray itself as a climate change governance leader (T. Rayner & Jordan, 2013; Wurzel & Connelly, 2011).

The fact that this thesis has pointed to shortcomings of polycentric governance (and its relationship to evaluation) demonstrates a need to both improve evaluation (see below in the section on policy recommendations), but also and crucially to continue developing the emerging theory of polycentric governance. While this thesis has attempted to address one influential line of criticism, namely the lack of empirical explorations of real-world examples of polycentric governance (see for example Jordan et al., 2015; Jordan, Huitema, van Asselt, & Forster, 2018; Morrison, 2017), additional lines of criticism of polycentric governance include its insufficient theorization (e.g. Carlisle & Gruby, 2017), meaning that many of the propositions or attributes of polycentric governance had not been fully developed, thus making it difficult to assess the performance of polycentric governance systems. Here, again, this thesis has made a key contribution by considering the role of policy evaluation as an important factor. Last, scholars have admonished the lack of attention to power dynamics in literatures on polycentric governance (e.g. Morrison et al., 2017;

Singleton, 2017). The argument goes that much of existing debates on (polycentric) governance often assume relatively equally powerful actors, which is often not the case (see Partzsch, 2017). There is ample room to explore these dynamics further in the future, especially with a view to the relationship of policy evaluation and power/control, which is increasingly drawing interest (see Duffy, 2017).

9.4 Policy recommendations

This thesis documents a wealth of climate policy evaluations between 1997 and 2014. But it also identifies a lack of sustained effort by actors to systematically bring that knowledge together in the form of databases or other knowledge sharing platforms. Therefore, the first recommendation is to create and maintain a publicly accessible climate policy evaluation database, which incorporates *both* formal and informal evaluations from the EU and if possible beyond. Network-managing actors such as the European Commission or the European Environment Agency may be candidates for this crucial task, but other informal (i.e. non-state) actors may also in principle step up. A (still limited) but good example of how this may look is the 'Forschungsradar' database with evaluations on the German Energiewende.[80] Based on the empirical insights from this thesis, doing so would be a starting point for formal and/or informal actors to conduct meta analyses in order to draw important insights from the well over 600 evaluations identified since the early 1990s, plus any additional ones published since 2014. A similar question applies to the international level – who can oversee and systematize the evaluations emerging from the Paris Agreement review and transparency processes? Would the UNFCCC play an important role as a network manager, or would non-state actors step up their efforts? The categories developed in the novel coding scheme in this thesis could serve as a way to assess the quality of the evaluations in the database, for example on the extent to which they contextualize their findings or provide translations into other languages.

[80] http://www.forschungsradar.de

For formal and informal actors in particular governance centres, this thesis recommends to use evaluation more actively in order to look beyond one's own immediate horizon. But doing so should incorporate paying attention to the context from which insights from other governance centres emerged and assess to what extent contextual similarities allow for lesson-drawing (or not). This is to avoid applying generic policy lessons without much knowledge of how these lessons came about in the first place. Thus, rather than working from the notion of 'one policy, one evaluation,' actors in any one governance centre would start from the notion of 'one policy, many evaluations,' and thereby acknowledge the growing plurality of evaluations and their corresponding insights. Analogously, the findings of this thesis suggest that rather than working from the assumption of 'one evaluation, one (intended) user,' the polycentric perspective would suggest 'one evaluation, many users.' But a crucial question is of course who would step up to overcome the collective action problem that invariably emerges when there are many different potential users who do not directly contribute to the evaluation?

The third recommendation is that while policy evaluation may facilitate polycentric climate governance, in the context of the EU, evaluation is certainly no panacea to address climate change and should thus not be viewed or understood as such. From the (arguably normative) perspective of polycentric climate governance in the EU, climate policy evaluation exhibits some, but not necessarily all the characteristics that would allow it to contribute strongly to the emergence of a polycentric governance system. For example, current climate policy evaluation in the EU by and large neglects certain sub-sectors (such as waste or agriculture), and many evaluations do not contextualize enough in order to enable lesson drawing in the way that this thesis proposes. While it has become fashionable to look to the EU and its constituent parts to derive policy lessons for other systems (Benson & Lorenzoni, 2014; Wyns, 2015), this thesis tells a more cautionary tale of the functioning of polycentric climate governance in the EU than others have done before (Cole, 2015; E. Ostrom, 2010c; see T. Rayner & Jordan, 2013) because some of its elements – notably evaluation – do not fully meet the expectations advanced by polycentric governance theory.

9.5 New research priorities

This thesis has laid a new foundation for future research to further the relationship between polycentric governance and policy evaluation. Based on the theoretical discussions and the empirical evidence, this thesis proposes five priority areas for further work: assessing causality; exploring evaluation use in polycentric systems; exploring additional aspects relevant for polycentric governance; linkages with other actors and practices; and studying other policy areas beyond climate change. This section discusses each area in turn.

The first area for future research involves exploring how and why the evaluation patterns and characteristics emerge in the first place. For example, what factors stimulate or inhibit self-organization in evaluation? Is the smaller number of informal evaluations due to a lower willingness to evaluate by informal actors, a lack of adequate resources, or perhaps even a lack of sufficient openness of the polycentric system? This thesis has assumed funding as an important impetus for evaluation, but how does its influence fare in comparison to other factors? Are there too few openings for informal actors to become involved in climate governance in the EU vis-à-vis evaluation? Recall that Vincent Ostrom had stressed that polycentric governance systems need to be sufficiently open to new actors in order to function (V. Ostrom, 1999a; V. Ostrom, 1999b). Similarly, why do we observe the strong congruence between the locations of the evaluation funders, the evaluators, as well as the focus of the climate policy under evaluation? Furthermore, researchers should explore what drives the correlations between self-organization and context, meaning that informal evaluations contextualize their findings more than formal evaluations on various dimensions? Galaz et al. (2012) argue that information sharing within polycentric governance system may eventually lead to stronger forms of collaboration, such as joint projects. But can such developments be empirically detected as a function of evaluation? This thesis has laid the groundwork to explore such causal mechanisms in climate governance in the EU and potentially elsewhere. Detailed process tracing and qualitative interview work with evaluation funders, evaluators and (potential) users could shed further light on these important questions and could complement the initial and thus partial perspective of this thesis (see Chapter 4).

Working towards more causal explanations could also involve considering alternatives to the differentiation between formal and informal evaluation based on evaluation funding. Alternatives include considering the nature of the actor that conducts the evaluation (for a more detailed discussion of this aspect, see Chapter 8), looking at the presence of legal requirements (see for example Wirths, Rosser, Horber-Papazian, & Mader, 2017), or measuring self-organization on a scale rather than with binary codes. For example, one could assess the degree to which an actor depends on public money, or has independence from policy-makers. Doing so would be enormously challenging given that these data would have to be generated through surveys or interviews—a notoriously difficult exercise, as Löwenbein (2008) experienced when he received a survey response rate of just 5% among professional evaluators in Germany—but an endeavour that could ultimately explain the finding that the formal-informal distinction also appears to blur in practice. For example, many Brussels-based interest groups receive significant funding from the EU institutions (see Greenwood, 2011). The empirical difficulty to categorize evaluations as either 'formal' or 'informal' necessitates exploring alternative categorizations and their consequences for evaluation production and use. Especially the principal-agent relationships, which have remained outside the scope of this thesis (see Pleger & Sager, 2016), are ripe for further assessment, including how they relate to polycentric governance arrangements. Exploring such causal mechanisms would work towards a deeper understanding of the *origins* of polycentric climate governance, and it could also enable *prediction* of future governance outcomes under certain conditions.

The second priority for future research involves exploring actual uses of evaluation knowledge in polycentric governance systems. It is a challenging, but nevertheless highly pertinent task to study whether the insights and recommendations from the evaluations actually translate into policy use and thus change (e.g. Johnson et al., 2009; van Voorst & Zwaan, 2018), while keeping in mind the full range of scholarship on knowledge utilization (see for example Rich, 1991; Rich, 1997). While the current thesis has probed interactions between governance centres directly through evaluation, as well as the potential for interaction as a function of evaluation characteristics, the next logical step involves investigating whether certain evaluation characteristics – such as for example the presence of executive summaries,

translations, or recommendations – actually translate into greater use (or its improvement). More interpretivist research approaches may also consider how different actors construct the concept of 'evaluation use,' and how these constructions impact on the evaluation process more generally. This could also link with an assessment of the various discourses that connect with evaluation (see Duffy, 2017). Research on evaluation use could thus significantly extend our understanding of interactions in polycentric governance systems and therefore ultimately allow much more definitive assessments of their effectiveness.

The third priority area involves exploring additional aspects that have gained increasing prominence in more recent debates on polycentric governance. Since the work on this thesis began in late 2013, Jordan et al. (2018) have for example identified five core theoretical 'propositions' related to polycentric governance theory, namely 'local action', 'mutual adjustment', 'experimentation', 'trust', and 'overarching rules.' While this thesis has explored some of these propositions with particular relevance to policy evaluation in detail, others have a theoretically much wider scope, but would equally benefit from much closer investigation with a view to evaluation. These include notably the roles of experimentation, trust and overarching rules. To date, little has, for example, been said about the link between governance experimentation and evaluation (see Bernstein & Hoffmann, 2018), and there is a dearth of work on potential links between evaluation and trust, as well as how the concepts of policy evaluation and overarching rules may hang together. Likewise, evaluation may have to play an important role in generating trust, which Dorsch and Flachsland (2017) highlight as one of the key ingredients of governance systems. And policy evaluation may help build trust – recall Chelimsky (2006, p. 54), who writes that in democracies "[i]t is distrust, once again, that generates the deepest constituency for evaluation." Furthermore, Chapter 2 briefly discussed questions on the accountability and legitimacy of polycentric governance arrangements. There is now an emerging debate on these concepts with a view to polycentric governance (Bäckstrand et al., 2018), and it would be useful to explicitly connect the thinking on evaluation in polycentric governance with these debates in order to widen the thinking on roles for evaluation in polycentric governance. For example, the fact that there is a significant number of informal evaluations indicates that policy evaluation could support horizontal accountability mechanisms in polycentric arrangements and

thus point to the emergence of new accountability chains that may better fit the nature of polycentric systems (see Chapter 2). In other words, further explorations of the *effects* of evaluation in polycentric governance systems would be helpful – including potentially the relationship between evaluation and control and/or transformation (see Duffy, 2017) in polycentric settings.

The fourth priority includes studying various forms of linkages between different evaluation actors in polycentric governance, where the debate is advancing quickly, but where evaluation has been little discussed (see Pattberg, Chan, Sanderink, & Widerberg, 2018). For instance, while there has certainly been increasing 'supply' and professionalization of evaluation activities around Europe (Brandt, 2009; Summa & Toulemonde, 2002), particularly through the establishment of the European Evaluation Society[81] and many national evaluation organizations (see also Duscha et al., 2009; Pollitt, 1998), we know too little about the extent to which these activities in organized evaluation societies correspond with policy evaluation practices in the European Commission, in other European institutions, in the Member States, as well as with the multiple informal actors that this thesis has identified – and thus potentially misses important interactions within the evaluation community. For example, evidence from Germany suggests that the evaluators that the state uses to conduct evaluations do not necessarily interact with evaluation societies (Duscha et al., 2009). The nature of these interactions remains to be explored in the case of climate change with the aim of working towards a fuller understanding of policy evaluation in Europe.

Linkages also matter with a view to other—and related—forms of policy assessment. For example, in the area of climate change governance, there is a huge gap in exploring how (*ex-ante*) impact assessments (see Radaelli, 2010; Turnpenny et al., 2016) connect with (*ex-post*) policy evaluations, including in the context of polycentric governance. For example, as Fritsch and colleagues (2013, p. 450) propose, a key way to assess the quality of *ex-ante* assessments would be to compare their estimates of policy effects with the findings of *ex-post* evaluations. The current

[81] http://www.europeanevaluation.org/

database of *ex-post* policy evaluations is a vital building block and indeed pre-condition for this type of research in the climate policy domain.

The final type of future work involves extending the theoretical and empirical insights from this thesis to other policy areas in order to assess whether the characteristics of evaluation in the climate sector – such as limited contextualization and potential for interaction, but also differences between formal and informal evaluations – hold in other cases. An area for potential comparative work may be ocean governance, which is a global issue where scholars have already explored aspects of polycentric governance (e.g., Galaz et al., 2012; Morrison, 2017). Comparison may help to identify whether evaluation amounts to a necessary condition for polycentric governance.

In closing, this thesis has thus not only covered important theoretical and empirical ground with a view to evaluation in polycentric (climate) governance, but it has also opened up numerous avenues for future research. In doing so, it contributes to a collective endeavour to address climate change and therefore work towards the successful implementation of polycentric climate governance through the Paris Agreement and potentially well beyond.

References

Abbott, K. W. (2011). The transnational regime complex for climate change. *Environment & Planning C: Government & Policy, 30*, 571-590.

AEA, ECOFYS, Fraunhofer, & ICCS. (2009). *Quantification of the effects on greenhouse gas emissions of policies and measures: Final Report.* (No. ENV.C.1/SER/2007/0019). Brussels: European Commission.

Albaek, E. (1995). Between knowledge and power: Utilization of social science in public policy making. *Policy Sciences, 28*(1), 79-100.

Aldy, J. E. (2014). The crucial role of policy surveillance in international climate policy. *Climatic Change, 126*(3-4), 279-292.

Aldy, J. E., & Pizer, W. A. (2014). *Comparability of effort in international climate policy architecture.* (No. HKS Working Paper No. RWP14-006). Boston: Harvard University.

Aldy, J. E., & Pizer, W. A. (2015). Alternative metrics for comparing domestic climate change mitigation efforts and the emerging international climate policy architecture. *Review of Environmental Economics and Policy, 10*(1), 3-24.

Aldy, J. E., Pizer, W. A., & Akimoto, K. (2017). Comparing emissions mitigation efforts across countries. *Climate Policy, 17*(4), 501-515.

Aligica, P. D. (2014). *Institutional diversity and political economy: The Ostroms and beyond*. Oxford; New York: Oxford University Press.

Aligica, P. D., & Sabetti, F. (2014a). The Ostroms' research program for the study of institutions and governance: Theoretical and epistemic foundations. In F.

Sabetti, & P. D. Aligica (Eds.), *Choice, rules and collective action: The Ostroms on the study of institutions and governance* (pp. 1-43). Colchester: ECPR Press.

Aligica, P. D., & Sabetti, F. (2014b). The collective action theory path to contextual analysis. *Journal of Natural Resources Policy Research, 6*(4), 253-258.

Alkin, M. C., & Christie, C. A. (2004). An evaluation theory tree. In M. C. Alkin (Ed.), *Evaluation roots: Tracing theorists' views and influences* (pp. 12-65). Thousand Oaks: Sage.

Austen-Smith, D., & Banks, J. S. (1998). Social choice theory, game theory, and positive political theory. *Annual Review of Political Science, 1*(1), 259-287.

Bäckstrand, K., Zelli, F., & Schleifer, P. (2018). Legitimacy and accountability in polycentric climate governance. In A. J. Jordan, D. Huitema, H. van Asselt & J. Forster (Eds.), *Governing climate change: Polycentricity in action?* (pp. 338-356). Cambridge: Cambridge University Press.

Bang, G., Victor, D. G., & Andresen, S. (2017). California's cap-and-trade system: Diffusion and lessons. *Global Environmental Politics, 17*(3), 12-30.

Batterbury, S. C. (2006). Principles and purposes of European Union cohesion policy evaluation. *Regional Studies, 40*(02), 179-188.

Bauböck, R. (2008). Normative political theory and empirical research. In D. Della Porta, & M. Keating (Eds.), *Approaches and methodologies in the social sciences* (pp. 40-60). Cambridge: Cambridge University Press.

Bemelmans-Videc, M. L. (1995). Evaluation in Europe and a new professional association: The EES. *Knowledge, Technology & Policy, 8*(3), 3-7.

Benjamin, R. (1982). The historical nature of social-scientific knowledge: The case of comparative political inquiry. In E. Ostrom (Ed.), *Strategies of political inquiry* (pp. 69-98). Beverly Hills: Sage Publications.

Bennett, C. J., & Howlett, M. (1992). The lessons of learning: Reconciling theories of policy learning and policy change. *Policy Sciences, 25*(3), 275-294.

Benson, D., & Jordan, A. J. (2011). What have we learned from policy transfer research? Dolowitz and Marsh revisited. *Political Studies Review, 9*(3), 366-378.

Benson, D., & Lorenzoni, I. (2014). Examining the scope for national lesson-drawing on climate governance. *The Political Quarterly, 85*(2), 202-211.

Bernstein, S., & Hoffmann, M. (2018). The politics of decarbonization and the catalytic impact of subnational climate experiments. *Policy Sciences,* doi:https://doi.org/10.1007/s11077-018-9314-8

Beywl, W., Fabian, C., & Widmer, T. (2009). *Evaluation: Ein systematisches Handbuch*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Bodansky, D. (1993). United Nations Framework Convention on Climate Change: A commentary. *Yale Journal of International Law, 18*, 451-558.

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

Bovens, M., Hart, P., & Kuipers, S. (2006). The politics of policy evaluation. In M. Moran, M. Rein & R. E. Goodin (Eds.), *The Oxford handbook of public policy* (pp. 319-335). Oxford; New York: Oxford University Press.

Bovens, M. (2007). New forms of accountability and EU-governance. *Comparative European Politics, 5*(1), 104-120.

Brandt, T. (2009). *Evaluation in Deutschland: Professionalisierungsstand und -perspektiven*. Münster; München: Waxmann.

Bressers, N., Twist, M., & Heuvelhof, E. (2013). Exploring the temporal dimension in policy evaluation studies. *Policy Sciences, 46*(1), 23-37.

Bugdahn, S. (2008). Travelling to Brussels via Aarhus: Can transnational NGO networks impact on EU policy? *Journal of European Public Policy, 15*(4), 588-606.

Bulkeley, H., Andonova, L., Betsill, M. M., Compagnon, D., Hale, T., Hoffmann, M. J., . . . VanDeveer, S. D. (2014). *Transnational climate change governance*. New York: Cambridge University Press.

Bundi, P. (2016). What do we know about the demand for evaluation? Insights from the parliamentary arena. *American Journal of Evaluation, 37*(4), 522-541.

Burnham, P. (2009). State. In I. McLean, & A. McMillan (Eds.), *The concise Oxford dictionary of politics* (3rd ed.,). Oxford: Oxford University Press.

Bussmann, W. (2005). Typen und Terminologie von Evaluationsklauseln. *LeGes, 16*(1), 97-102.

Bussmann, W. (2008). The emergence of evaluation in Switzerland. *Evaluation, 14*(4), 499-506.

Carlisle, K., & Gruby, R. L. (2017). Polycentric systems of governance: a theoretical model for the commons. *Policy Studies Journal,* doi:10.1111/psj.12212

Chelimsky, E. (2006). The purposes of evaluation in a democratic society. In I. Shaw, J. Greene & M. Mark (Eds.), *The SAGE handbook of evaluation* (pp. 33-55). London: Sage Publications.

Chelimsky, E. (2009). Integrating evaluation units into the political environment of government: The role of evaluation policy. *New Directions for Evaluation, 123*, 51-66.

Christiansen, T., & Neuhold, C. (2013). Informal politics in the EU. *Journal of Common Market Studies, 51*(6), 1196-1206.

Christie, C. A. (2003). What guides evaluation? A study of how evaluation practice maps onto evaluation theory. *New Directions for Evaluation, 97*(97), 7-36.

Cole, D. H. (2015). Advantages of a polycentric approach to climate change policy. *Nature Climate Change, 5*, 114-118.

Cole, D. H. (2011). From global to polycentric climate governance. *Climate Law, 2*(3), 395-413.

Conner, R. F., Fitzpatrick, J. L., & Rog, D. J. (2012). A first step forward: Context assessment. *New Directions for Evaluation, 2012*(135), 89-105.

COWI. (2013). Evaluation of the European Environment Agency. Retrieved from http://www.eea.europa.eu/about-us/governance/eea-evaluations/eea-evaluation-2013

Crabbé, A., & Leroy, P. (2008). *The handbook of environmental policy evaluation*. London; Sterling, VA: Earthscan.

Dabinett, G., & Richardson, T. (1999). The European spatial approach: The role of power and knowledge in strategic planning and policy evaluation. *Evaluation, 5*(2), 220-236.

Dahler-Larsen, P. (2011). Organizing knowledge: Evidence and the construction of evaluative information systems. In R. C. Rist, & N. Stame (Eds.), *From studies to streams: managing evaluative systems* (pp. 65-80). New Brunswick: Transaction Publishers.

De Winter, J. C. F., & Dodou, D. (2010). Five-point Likert items: t test versus Mann-Whitney-Wilcoxon. *Practical Assessment, Research & Evaluation, 15*(11), 1-12.

Derlien, H. U. (1990). Program evaluation in the Federal Republic of Germany. In R. C. Rist (Ed.), *Program Evaluation and the Management of Government:*

Patterns and Prospects Across Eight Nations (pp. 37-51). New Brunswick: Transaction Publishers.

Derlien, H. U. (2002). Policy evaluation in Germany: Institutional continuation and sectoral activation. In J. E. Furubo, R. C. Rist & R. Sandahl (Eds.), *International atlas of evaluation* (pp. 77-91). New Brunswick: Transaction Publishers.

Diaz-Puente, J. M., Yaguee, J. L., & Afonso, A. (2008). Building evaluation capacity in Spain - A case study of rural development and empowerment in the European Union. *Evaluation Review, 32*(5), 478-506.

D'Lane, C., Love, T. P., & Sell, J. (2012). Developing and assessing intercoder reliability in studies of group interaction. *Sociological Methodology, 42*(1), 348-364.

Doll, C., Eichhammer, W., Fleiter, T., Jochem, E., Köhler, J., Peters, A., . . . Ziesing, H. J. (2012). *Ermittlung der Klimaschutzwirkung des Integrierten Energie- und Klimaschutzprogramms der Bundesregierung IEKP und Vorschlag für ein Konzept zur kontinuierlichen Überprüfung der Klimaschutzwirkung des IEKP - Arbeitspaket 2: Entwicklung eines Monitoringkonzepts für das Integrierte Energie- und Klimaschutzprogramm (IEKP)*. Dessau-Roßlau: Umweltbundesamt.

Donaldson, S. I., & Lipsey, M. W. (2006). Roles for theory in contemporary evaluation practice: Developing practical knowledge. In I. F. Shaw, J. C. Greene & M. M. Mark (Eds.), *The SAGE handbook of evaluation* (pp. 56-75). London: SAGE Publications.

Dorsch, M. J., & Flachsland, C. (2017). A polycentric approach to global climate governance. *Global Environmental Politics, 17*(2), 45-64.

Duffy, D. N. (2017). *Evaluation and governing in the 21st century: Disciplinary measures, transformative possibilities*. London: Palgrave Macmillan.

Dunleavy, P. (2003). *Authoring a PhD: How to plan, draft, write and finish a doctoral thesis or dissertation*. Basingstoke: Palgrave Macmillan.

Dupont, C., & Oberthür, S. (2015). Decarbonization in the EU: Setting the scene. In C. Dupont, & S. Oberthür (Eds.), *Decarbonization in the European Union: Internal policies and external strategies* (pp. 1-24). Basingstoke: Palgrave McMillan.

Dupuis, J., & Biesbroek, R. (2013). Comparing apples and oranges: The dependent variable problem in comparing and evaluating climate change adaptation policies. *Global Environmental Change,*

Duscha, M., Klemisch, H., & Meyer, W. (2009). Umweltevaluation in Deutschland - Entwicklungstrends mit Fokus auf dem Energiesektor. In T. Widmer, W. Beywl & C. Fabian (Eds.), *Evaluation: Ein systematisches Handbuch* (pp. 203-212). Wiesbaden: VS Verlag für Sozialwissenschaften.

Eckstein, H. (2000). Case study and theory in political science. In R. Gomm, M. Hammersley & P. Foster (Eds.), *Case study method: key issues, key texts* (pp. 119-164). Thousand Oaks, CA: Sage.

European Commission. (1996). *Evaluation: Concrete steps towards best practice across the Commission.* (No. SEC 96/659 final).European Commission.

European Commission. (2002). Evaluation standards. Retrieved from http://ec.europa.eu/smart-regulation/evaluation/docs/standards_c_2002_5267_final_en.pdf

European Commission. (2010). *Smart Regulation in the European Union*. Brussels: European Commission.

European Commission. (2013). Strengthening the foundations of smart regulation - improving evaluation. Retrieved from http://ec.europa.eu/smart-regulation/docs/com_2013_686_en.pdf

European Environment Agency. (2001). *Reporting on environmental measures: Are we being effective?* (No. 25). Luxembourg: Office for Official Publications of the European Communities.

European Environment Agency. (2016). *Environment and Climate Policy Evaluation*. Copenhagen: European Environment Agency.

Feldman, D. L., & Wilt, C. A. (1996). Evaluating the implementation of state-level global climate change programs. *Journal of Environment and Development*, *5*(1), 46-72.

Fetterman, D. M., & Wandersman, A. (2005). *Empowerment evaluation principles in practice*. New York: Guilford Press.

Fischer, F. (2006). *Evaluating public policy*. Mason: Cengage Learning.

Fitzpatrick, J. L. (2012). An introduction to context and its role in evaluation practice. *New Directions for Evaluation, 2012*(135), 7-24.

Flyvbjerg, B. (2006). Five misunderstandings about case-study research. *Qualitative Inquiry, 12*(2), 219-245.

Fournier, D. M. (2005). Evaluation. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 139-140). Thousand Oaks: Sage.

Frankfort-Nachmias, C., & Nachmias, D. (1996). Ethics in social science research. *Reserach methods in the social sciences* (2nd ed., pp. 75-96). London: Arnold.

Fransen, T., & Cronin, C. (2013). A critical decade for climate policy: Tools and initiatives to track our progress. Retrieved from

http://www.wri.org/sites/default/files/pdf/critical_decade_for_climate_policy_tools_and_initiatives_to_track_our_progress.pdf

Fritsch, O., Radaelli, C. M., Schrefler, L., & Renda, A. (2013). Comparing the content of regulatory impact assessments in the UK and the EU. *Public Money & Management, 33*(6), 445-452.

Furlong, P., & Marsh, D. (2010). A skin not a sweater: Ontology and epistemology in political science. In D. Marsh, & G. Stoker (Eds.), *Theory and methods in political science* (pp. 184-211). Basingstoke: Palgrave Macmillan.

Furubo, J. E., Rist, R. C., & Sandahl, R. (2002). *International atlas of evaluation*. New Brunswick: Transaction Publishers.

Galaz, V., Crona, B., Österblom, H., Olsson, P., & Folke, C. (2012). Polycentric systems and interacting planetary boundaries—Emerging governance of climate change–ocean acidification–marine biodiversity. *Ecological Economics, 81*, 21-32.

Gallopín, G. C. (1996). Environmental and sustainability indicators and the concept of situational indicators: A systems approach. *Environmental Modeling & Assessment, 1*(3), 101-117.

George, A. L., & Bennett, A. (2005). *Case studies and theory development in the social sciences*. Cambridge, Mass: MIT Press.

Goodin, R. E. (1996). *The theory of institutional design*. Cambridge [England]; New York, NY, USA: Cambridge University Press.

Gore, T., & Wells, P. (2009). Governance and evaluation: The case of EU regional policy horizontal priorities. *Evaluation and Program Planning, 32*(2), 158-167.

Gravey, V. (2016). *Does the European Union have a reverse gear? Environmental policy dismantling, 1992-2014.* (Unpublished PhD). University of East Anglia,

Norwich. Retrieved from https://ueaeprints.uea.ac.uk/59419/1/ThesisVGravey.pdf

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

Greene, J. C. (2005). Context. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 82-84). Thousand Oaks: Sage.

Greene, J. C. (1997). Evaluation as advocacy. *American Journal of Evaluation, 18*(1), 25-35.

Greenwood, J. (2011). *Interest representation in the European Union* (3rd ed.). New York: Palgrave Macmillan.

Guba, E. G., & Lincoln, Y. S. (1981). *Effective evaluation*. San Francisco: Josey-Bass Limited.

Guba, E. G., & Lincoln, Y. S. (1989). *Fourth generation evaluation*. Newbury Park: Sage Publications.

Gudmundsson, H. (2003). The policy use of environmental indicators—learning from evaluation research. *The Journal of Transdisciplinary Environmental Studies, 2*(2), 1-12.

Guha-Khasnobis, B., Kanbur, R., & Ostrom, E. (2006). Beyond formality and informality. In B. Guha-Khasnobis, R. Kanbur & E. Ostrom (Eds.), *Linking the formal and informal economy: Concepts and policies* (pp. 1-18). Oxford; New York: Oxford University Press.

Gupta, J. (2008). Global change: Analyzing scale and scaling in environmental governance. In O. R. Young, L. A. King & H. Schroeder (Eds.), *Institutions and*

environmental change: principal findings, applications, and research frontiers (pp. 225-258). Cambridge, Mass.: MIT Press.

Haas, P. M. (1992). Introduction: Epistemic communities and international policy coordination. *International Organization, 46*(1), 1-35.

Haas, P. (2004). When does power listen to truth? A constructivist approach to the policy process. *Journal of European Public Policy, 11*(4), 569-592.

Haigh, N. (1996). Climate change policies and politics in the European Community. In T. O'Riordan, & J. Jäger (Eds.), *Politics of climate change: A European perspective* (pp. 155-185). New York: Routledge.

Hale, T., & Roger, C. (2013). Orchestration and transnational climate governance. *The Review of International Organizations, 9*, 59-82.

Hanberger, A. (2012). Framework for exploring the interplay of governance and evaluation. *Scandinavian Journal of Public Administration, 16*(3), 9-27.

Hardin, G. (1968). The tragedy of the commons. 1968. *Science, 162*(3859), 1243-1248.

Haug, C. (2015). *Unpacking learning: Conceptualising and measuring the effects of two policy exercises on climate governance.* (Unpublished PhD). Vrije Universiteit, Amsterdam. Retrieved from http://dare.ubvu.vu.nl/bitstream/handle/1871/53559/complete%20dissertation.pdf?sequence=1

Haug, C., Rayner, T., Jordan, A. J., Hildingsson, R., Stripple, J., Monni, S., . . . Berkhout, F. (2010). Navigating the dilemmas of climate policy in Europe: Evidence from policy evaluation studies. *Climatic Change, 101*(3-4), 427-445.

Hay, C. (2002). What's political about political science? *Political analysis* (pp. 59-88). Houndmills, Basingstoke, Hampshire; New York: Palgrave.

Hayward, R. J., Kay, J., Lee, A., Page, E. C., Patel, N., Payne, H., . . . Valyraki, A. (2013). Evaluation under contract: Government pressure and the production of policy research. *Public Administration, 92*(1), 224-239.

Helmke, G., & Levitsky, S. (2004). Informal institutions and comparative politics: A research agenda. *Perspectives on Politics, 2*(04), 725-740.

Hertting, N., & Vedung, E. (2012). Purposes and criteria in network governance evaluation: How far does standard evaluation vocabulary takes us? *Evaluation, 18*(1), 27-46.

Hildén, M., Jordan, A. J., & Rayner, T. (2014). Climate policy innovation: developing an evaluation perspective. *Environmental Politics, 23*(5), 884-905.

Hildén, M. (2009). Time horizons in evaluating environmental policies. *New Directions for Evaluation, 2009*(122), 9-18.

Hildén, M. (2014). Evaluation, assessment, and policy innovation: Exploring the links in relation to emissions trading. *Environmental Politics, 23*(5), 839-859.

Hildén, M., Jordan, A., & Huitema, D. (2017). Special issue on experimentation for climate change solutions editorial: The search for climate change and sustainability solutions - The promise and the pitfalls of experimentation. *Journal of Cleaner Production, 169*, 1-7.

Hill, M., & Varone, F. (2017). *The public policy process*. Oxon: Routledge.

HM Treasury. (2003). The green book - Appraisal and evaluation in central government. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/220541/green_book_complete.pdf

HM Treasury. (2011). The magenta book: Guidance for evaluation. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/220542/magenta_book_combined.pdf

HM Treasury, & DECC. (2010). Valuation of energy use and greenhouse gas emissions for appraisal and evaluation. Retrieved from http://webarchive.nationalarchives.gov.uk/20110612080604/http://decc.gov.uk/assets/decc/statistics/analysis_group/122-valuationenergyuseggemissions.pdf

Hogwood, B. W., & Gunn, L. A. (1984). *Policy analysis for the real world*. Oxford: Oxford University Press.

Hojlund, S. (2015). Evaluation in the European Commission. *European Journal of Risk Regulation, 1*, 35-46.

House, E., & Howe, K. R. (1999). *Values in evaluation and social research*. Thousand Oaks, California: Sage.

Howlett, M. (2014). Why are policy innovations rare and so often negative? Blame avoidance and problem denial in climate change policy-making. *Global Environmental Change, 29*, 395-403.

Huitema, D., Jordan, A. J., Massey, E., Rayner, T., Van Asselt, H., Haug, C., . . . Stripple, J. (2011). The evaluation of climate policy: Theory and emerging practice in Europe. *Policy Sciences, 44*(2), 179-198.

Hulme, M., Neufeldt, H., Colyer, H. & Ritchie, A. (2009). Adaptation and mitigation strategies: Supporting European climate policy. The final report from the ADAM project. Retrieved from http://www.tyndall.ac.uk/sites/default/files/adam-final-report-revised-june-2009.html_.pdf

Hyvarinen, J. (1999). The European Community's Monitoring Mechanism for CO2 and other greenhouse gases: The Kyoto Protocol and other recent developments. *Review of European Community & International Environmental Law, 8*(2), 191-197.

IEEP, & EIPA. (2003). Evaluation of the European Environment Agency. Retrieved from http://ec.europa.eu/environment/pubs/pdf/eea_c_en.pdf

Jackson, R. B., LeQuéré, C., Andrew, R. M., Canadell, J. G., Peters, G. P., Roy, J., & Wu, L. (2017). Warning signs for stabilizing global CO2 emissions. *Environmental Research Letters, 12*(11) doi:DOI: 10.1088/1748-9326/aa9662

Jacob, S., Speer, S., & Furubo, J. (2015). The institutionalization of evaluation matters: Updating the International Atlas of Evaluation 10 years later. *Evaluation, 21*(1), 6-31.

Jacobs, M. (2012). Climate policy: Deadline 2015. *Nature*, *481*(7380), 137-138.

Jenkins, B., & Gray, A. (1990). Policy evaluation in British government: From idealism to realism. In R. C. Rist (Ed.), *Program evaluation and the management of government: Patterns and prospects across eight nations* (pp. 53-70). New Brunswick: Transaction Publishers.

Johnson, K., Greenseid, L. O., Toal, S. A., King, J. A., Lawrenz, F., & Volkov, B. (2009). Research on evaluation use: A review of the empirical literature from 1986 to 2005. *American Journal of Evaluation, 30*(3), 377-410.

Jordan, A. J. (2001). The European Union: An evolving system of multi-level governance... or government? *Policy & Politics, 29*(2), 193-208.

Jordan, A. J., & Huitema, D. (2014). Innovations in climate policy: The politics of invention, diffusion, and evaluation. *Environmental Politics, 23*(5), 715-734.

Jordan, A. J., Huitema, D., Hildén, M., van Asselt, H., Rayner, T., Schoenefeld, J. J., . . . Boasson, E. L. (2015). Emergence of polycentric climate governance and its future prospects. *Nature Climate Change, 5*, 977-982.

Jordan, A. J., Huitema, D., Schoenefeld, J. J., van Asselt, H., & Forster, J. (2018). Governing climate change polycentrically: setting the scene. In A. J. Jordan, D. Huitema, H. van Asselt & J. Forster (Eds.), *Governing climate change: polycentricity in action?* (pp. 3-25). Cambridge: Cambridge University Press.

Jordan, A. J., Huitema, D., Van Asselt, H., Rayner, T., & Berkhout, F. (2010). *Climate change policy in the European Union: Confronting the dilemmas of mitigation and adaptation?*. Cambridge; New York: Cambridge University Press.

Jordan, A. J., Huitema, D., van Asselt, H., & Forster, J. (Eds.). (2018). *Governing climate change: Polycentricity in action?*. Cambridge: Cambridge University Press.

Jordan, A. J., & Schout, A. (2006). *The coordination of the European Union: exploring the capacities of networked governance*. Oxford; New York: Oxford University Press.

Jordan, A. J., Van Asselt, H., Berkhout, F., Huitema, D., & Rayner, T. (2012). Climate change policy in the European Union: Understanding the paradoxes of multi-level governing. *Global Environmental Politics, 12*(2), 43-66.

Kaufmann, P., & Wangler, L. U. (2014). Evaluation in der Umweltpolitik am Beispiel Klimapolitik. In W. Böttcher, C. Kerlen, P. Maats, O. Schwab & S. Sheikh (Eds.), *Evaluation in Deutschland und Österreich: Stand und Entwicklungsperspektiven in den Arbeitsfeldern der DeGEval-Gesellschaft für Evaluation* (pp. 159-170). Münster: Waxmann Verlag.

Keohane, R. O., & Ostrom, E. (1994). *Local commons and global interdependence*. Thousand Oaks: Sage.

Kerr, A. (2007). Serendipity is not a strategy: The impact of national climate programmes on greenhouse-gas emissions. *Area, 39*(4), 418-430.

King, J. A. (2003). The challenge of studying evaluation theory. *New Directions for Evaluation, 2003*(97), 57-68.

Kleine, M. (2013). Informal governance in the European Union. *Journal of European Public Policy, 21*(2), 303-314.

Knaap, G. J., & Kim, T. J. (1998). *Environmental program evaluation: A primer*. Urbana: University of Illinois Press.

Kusek, J. Z., & Rist, R. C. (2005). Performance-based monitoring. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 300-306). Thousand Oaks: Sage.

Lange, E. (1983). Zur Entwicklung und Methodik der Evaluationsforschung in der Bundesrepublik Deutschland. *Zeitschrift Für Soziologie, 12*(3), 253-270.

Lascoumes, P., & Le Gales, P. (2007). Introduction: Understanding public policy through its instruments—from the nature of instruments to the sociology of public policy instrumentation. *Governance, 20*(1), 1-21.

Lee, E., Su Jung, C., & Lee, M. (2014). The potential role of boundary organizations in the climate regime. *Environmental Science & Policy, 36*, 24-36.

Lehtonen, M. (2015). Indicators: Tools for informing, monitoring or controlling? In A. Jordan, & J. Turnpenny (Eds.), *The tools of policy formulation: Actors, capacities, venues and effects* (pp. 76-99). Cheltenham: Edward Elgar.

Levine, R. A. (1984). Programmevaluierung und Politikanalyse in Europa, USA und Kanada - ein Überblick. In G. M. Hellstem, & H. Wollmann (Eds.), *Handbuch zur Evaluierungsforschung* (pp. 94-133) Westdeutscher Verlag.

Lilliestam, J., Battaglini, A., Finlay, C., Fürstenwerth, D., Patt, A., Schellekens, G., & Schmidt, P. (2012). An alternative to a global climate deal may be unfolding before our eyes. *Climate and Development, 4*(1), 1-4.

Löwenbein, O. (2008). The evaluation market in Germany. *Journal of Multidisciplinary Evaluation, 5*(10), 78-88.

Lowi, T. J. (1972). Four systems of policy, politics, and choice. *Public Administration Review, 32*(4), 298-310.

Lukes, S. (2005). *Power: A radical view*. New York: Palgrave Macmillan.

Majone, G. (1989). *Evidence, argument and persuasion in the policy process*. New Haven: Yale University press.

Mangis, J. K. (1998). *The United States' sulfur dioxide emissions allowance program: An overview with emphasis of monitoring requirements and procedures and a summary report on US experience with environmental trading systems*. Copenhagen: European Environment Agency.

Mansbridge, J. J. (1999). Everyday talk in the deliberative system. In S. Macedo (Ed.), *Deliberative politics: essays on democracy and disagreement* (pp. 211-239). London: Oxford University Press.

Mansbridge, J. J. (2014). The role of the state in governing the commons. *Environmental Science and Policy, 36*, 8-10.

Marsh, D., & Sharman, J. C. (2009). Policy diffusion and policy transfer. *Policy Studies, 30*(3), 269-288.

Martens, M. (2010). Voice or loyalty? The evolution of the European Environment Agency (EEA). *Journal of Common Market Studies, 48*(4), 881-901.

Martin, R. (2001). Geography and public policy: The case of the missing agenda. *Progress in Human Geography, 25*(2), 189-210.

Mastenbroek, E., van Voorst, S., & Meuwese, A. (2016). Closing the regulatory cycle? A meta evaluation of ex-post legislative evaluations by the European Commission. *Journal of European Public Policy, 23*(9), 1329-1348.

McConnell, A. (2010). *Understanding policy success: Rethinking public policy*. Basingstoke: Palgrave Macmillan.

McGinnis, M., & Ostrom, E. (2008). Will lessons from small-scale social dilemmas scale up? In A. Biel, D. Eek, T. Gärling & M. Gustafsson (Eds.), *New issues and paradigms in research on social dilemmas* (pp. 189-211). New York: Springer.

McGinnis, M. D. (2011). Networks of adjacent action situations in polycentric governance. *Policy Studies Journal, 39*(1), 51-78.

McGinnis, M. D. (2016). *Polycentric governance in theory and practice: Dimensions of aspiration and practical limitations*. Bloomington: Paper presented at the Polycentricity Workshop, Indiana University, 14–17 December 2015.

McGinnis, M. D., & Ostrom, E. (2012). Reflections on Vincent Ostrom, public administration, and polycentricity. *Public Administration Review, 72*(1), 15-25.

Mela, H., & Hildén, M. (2012). Evaluation of climate policies and measures in EU member states–examples and experiences from four sectors. Retrieved from https://helda.helsinki.fi/bitstream/handle/10138/38749/FE19_2012.pdf?sequence=1

Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Google Books Team, . . . Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science, 331*(6014), 176-182.

Mickwitz, P. (2013). Policy evaluation. In A. Jordan, & C. Adelle (Eds.), *Environmental policy in the EU: Actors, institutions and processes* (pp. 267-286). London; New York: Routledge.

Mickwitz, P. (2003). A framework for evaluating environmental policy instruments: Context and key concepts. *Evaluation, 9*(4), 415-436.

Mickwitz, P. (2006). *Environmental policy evaluation: Concepts and practice*. Helsinki: The Finnish Society of Sciences and Letters.

Mitchell, S. D. (2009). *Unsimple truths: science, complexity, and policy*. Chicago: University of Chicago Press.

Morrison, T. H., Adger, W. N., Brown, K., Lemos, M. C., Huitema, D., & Hughes, T. P. (2017). Mitigation and adaptation in polycentric systems: Sources of power in the pursuit of collective goals. *Wiley Interdisciplinary Reviews: Climate Change*, doi:10.1002/wcc.479

Morrison, T. H. (2017). Evolving polycentric governance of the Great Barrier Reef. *Proceedings of the National Academy of Sciences of the United States of America, 114*(15), E3013-E3021.

Neslen, A. (2011). EU faces down tar sands industry. Retrieved from http://www.euractiv.com/climate-environment/eu-faces-tar-sands-industry-news-508140

Nilsson, M., Jordan, A. J., Turnpenny, J., Hertin, J., Nykvist, B., & Russel, D. (2008). The use and non-use of policy appraisal tools in public policy making: An analysis of three European countries and the European Union. *Policy Sciences, 41*(4), 335-355.

Oberthür, S. (2016). Where to go from Paris? The European Union in climate geopolitics. *Global Affairs, 2*(2), 119-130.

OECD-DAC. (2002). *Glossary of key terms in evaluation and results based management*. Paris: DAC Network on Development Evaluation, OECD.

Öko-Institut, Cambridge Economics, AMEC, Harmelink Consulting, & TNO. (2012). *Ex-post quantification of the effects and costs of policies and measures.* (No. CLIMA.A.3/SER/2010/0005). Berlin: Öko-Institut.

Ostrom, E. (2007). Institutional rational choice: An assessment of the institutional analysis and development framework. In P. A. Sabatier (Ed.), *Theories of the policy process* (2nd ed., pp. 21-64). Boulder, Colorado: Westview Press.

Ostrom, E. (Ed.). (1982). *Strategies of political inquiry*. Beverly Hills: Sage.

Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action*. Cambridge: Cambridge University Press.

Ostrom, E. (1999). Coping with tragedies of the commons. *Annual Review of Political Science, 2*(1), 493-535.

Ostrom, E. (2005). *Understanding institutional diversity*. Princeton, NJ: Princeton University Press.

Ostrom, E. (2006). Converting threats into opportunities. *Political Science & Politics, 39*(01), 3-12.

Ostrom, E. (2010a). Beyond markets and states: Polycentric governance of complex economic systems. *Transnational Corporations Review, 2*(2), 1-12.

Ostrom, E. (2010b). A multi-scale approach to coping with climate change and other collective action problems. *Solutions, 1*(2), 27-36.

Ostrom, E. (2010c). Polycentric systems for coping with collective action and global environmental change. *Global Environmental Change, 20*(4), 550-557.

Ostrom, E. (2012). Nested externalities and polycentric institutions: Must we wait for global solutions to climate change before taking actions at other scales? *Economic Theory, 49*(2), 353-369.

Ostrom, E. (2014a). Beyond Positivism. In F. Sabetti, & P. Dragos Aligia (Eds.), *Choice, rules and collective action: The Ostroms on the study of institutions and governance* (pp. 213-225). Colchester: ECPR Press.

Ostrom, E. (2014b). A polycentric approach for coping with climate change. *Annals of Economics and Finance, 15*(1), 71-108.

Ostrom, E., & Nagendra, H. (2007). Tenure alone is not sufficient: Monitoring is essential. *Environmental Economics & Policy Studies, 8*(3), 175-199.

Ostrom, V. (1999a). Polycentricity (Part 1). In M. D. McGinnis (Ed.), *Polycentricity and local public economies: Readings from the Workshop in Political Theory and Policy Analysis* (pp. 52-74). Ann Arbor: University of Michigan Press.

Ostrom, V. (1999b). Polycentricity (Part 2). In M. D. McGinnis (Ed.), *Polycentricity and local public economies: Readings from the Workshop in Political Theory and Policy Analysis* (pp. 119-138). Ann Arbor: University of Michigan Press.

Ostrom, V., Tiebout, C. M., & Warren, R. (1961). The organization of government in metropolitan areas: A theoretical inquiry. *The American Political Science Review, 55*(4), 831-842.

Ostrom, E., Janssen, M. A., & Anderies, J. M. (2007). Going beyond panaceas. *Proceedings of the National Academy of Sciences of the United States of America, 104*(39), 15176-15178.

Owens, S., Rayner, T., & Bina, O. (2004). New agendas for appraisal: Reflections on theory, practice, and research. *Environment and Planning A, 36*(11), 1943-1960.

Parsons, W. (2007). Policy analysis in Britain. In F. Fischer, G. J. Miller & M. S. Sidney (Eds.), *Handbook of public policy analysis: Theory, politics, and methods* (pp. 537-552). London: Taylor & Francis.

Partzsch, L. (2017). ‘Power with’and ‘power to’in environmental politics and the transition to sustainability. *Environmental Politics*, *26*(2), 193-211.

Pattberg, P., Chan, S., Sanderink, L., & Widerberg, O. (2018). Linkages: Understanding their role in polycentric governance. In A. J. Jordan, D. Huitema, H. van Asselt & J. Forster (Eds.), *Governing climate change: Polycentricity in action?* (pp. 169-187). Cambridge: Cambridge University Press.

Patton, M. Q. (2008). *Utilization-focused evaluation*. Los Angeles: Sage Publications.

Pattyn, V., Van Voorst, S., Mastenbroek, E., & Dunlop, C. A. (2018). Policy evaluation in Europe. *The Palgrave handbook of public administration and management in Europe* (pp. 577-593). London: Palgrave Macmillan.

Pawson, R., & Tilley, N. (1997). *Realistic evaluation*. London; Thousand Oaks, Calif: Sage.

Peters, B. G. (1998). Managing horizontal government: The politics of co-ordination. *Public Administration*, *76*(2), 295-311.

Peterson, J., & Shackleton, M. (2012). *The institutions of the European Union*. Oxford: Oxford University Press.

Piketty, T. (2014). *Capital in the twenty-first century* (A. Goldhammer Trans.). Cambridge, MA: The Belknap Press of Harvard University Press.

Pleger, L., & Sager, F. (2016). Betterment, undermining, support and distortion: A heuristic model for the analysis of pressure on evaluators. *Evaluation and Program Planning*, doi:https://doi.org/10.1016/j.evalprogplan.2016.09.002

Pollitt, C. (1993). Occasional excursions: A brief history of policy evaluation in the UK. *Parliamentary Affairs*, *46*(3), 353-363.

Pollitt, C. (1998). Evaluation in Europe: Boom or bubble? *Evaluation*, *4*(2), 214-224.

Pollitt, C. (2008). *Time, policy, management: Governing with the past*. Oxford: Oxford University Press.

Pollitt, C. (2013). *Context in public policy and management: The missing link?*. Cheltenham: Edward Elgar Publishing.

Polverari, L., & Bachtler, J. (2004). Assessing the evidence: The evaluation of regional policy in Europe. *EoRPA Paper, 4*(5), Paper prepared for the the EPRC Regional Policy Research Consortium at Ross Priory, Loch Lomondside on 3-5 October 2004.

Poptcheva, E. M. (2013). Policy and legislative evaluation in the EU. Retrieved from http://www.europarl.europa.eu/eplibrary/Policy-and-legislative-evaluation-in-the-EU.pdf

Poteete, A. R., Janssen, M., & Ostrom, E. (2010). *Working together: Collective action, the commons, and multiple methods in practice*. Princeton: Princeton University Press.

Radaelli, C. M. (1995). The role of knowledge in the policy process. *Journal of European Public Policy, 2*(2), 159-183.

Radaelli, C. M. (2007). Whither better regulation for the Lisbon agenda? *Journal of European Public Policy, 14*(2), 190-207.

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

Radaelli, C. M., & Dente, B. (1996). Evaluation strategies and analysis of the policy process. *Evaluation, 2*(1), 51-66.

Raupach, M. R., Davis, S. J., Peters, G. P., Andrew, R. M., Canadell, J. G., Ciais, P., . . . Le Quere, C. (2014). Sharing a quota on cumulative carbon emissions. *Nature Climate Change, 4*(10), 873-879.

Rayner, J., Howlett, M., Wilson, J., Cashore, B., & Hoberg, G. (2001). Privileging the sub-sector: Critical sub-sectors and sectoral relationships in forest policy-making. *Forest Policy and Economics*, *2*(3-4), 319-332.

Rayner, T., & Jordan, A. J. (2013). The European Union: The polycentric climate policy leader? *Wiley Interdisciplinary Reviews: Climate Change, 4*(2), 75-90.

Rich, R. F. (1991). Knowledge creation, diffusion, and utilization: Perspectives of the founding editor of knowledge. *Science Communication, 12*(3), 319-337.

Rich, R. F. (1997). Measuring knowledge utilization: Processes and outcomes. *Knowledge and Policy, 10*(3), 11-24.

Risley, J. S. (2007). Evaluation activities in the United Kingdom. *Journal of MultiDisciplinary Evaluation, 1*(1), 77-80.

Rist, R. C., & Stame, N. (2011). *From studies to streams: Managing evaluative systems*. New Brunswick: Transaction Publishers.

Rog, D. J. (2012). When background becomes foreground: Toward context-sensitive evaluation practice. *New Directions for Evaluation, 2012*(135), 25-40.

Rose, R. (1991). What is lesson-drawing? *Journal of Public Policy, 11*(1), 3-30.

Rose, R. (1993). *Lesson-drawing in public policy: A guide to learning across time and space*. New Jersey: Chatham House.

Sager, F., Widmer, T., & Balthasar, A. (2017). Schlussfolgerungen: Evaluation als Teil des politischen Systems der Schweiz? In F. Sager, T. Widmer & A. Balthasar (Eds.), *Evaluation im politischen System der Schweiz* (pp. 313-330). Zürich: NZZ Libro.

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

Schoenefeld, J. J., Hildén, M., & Jordan, A. J. (2018). The challenges of monitoring national climate policy: Learning lessons from the EU. *Climate Policy, 18*(1), 118-128.

Schoenefeld, J. J., & Jordan, A. J. (2017). Governing policy evaluation? Towards a new typology. *Evaluation, 23*(3), 274-293.

Schoenefeld, J. J., & McCauley, M. R. (2016). Local is not always better: The impact of climate information on values, behavior and policy support. *Journal of Environmental Studies and Sciences, 6*(4), 724-732.

Schröter, D. C. (2007). Evaluation in Europe: An overview. *Journal of Multidisciplinary Evaluation, 1*(1), 68-76.

Scriven, M. (1981). *Evaluation thesaurus*. Newbury Park: Sage.

Segerholm, C. (2003). Researching evaluation in national (state) politics and administration: A critical approach. *American Journal of Evaluation, 24*(3), 353-372.

Singleton, B. E. (2017). What's missing from Ostrom? Combining design principles with the theory of sociocultural viability. *Environmental Politics, 26*(6), 994-1014.

Smith, R. M. (2004). Reconnecting political theory to empirical inquiry, or, a return to the cave? In E. D. Mansfield, & R. Sisson (Eds.), *The evolution of political knowledge: Theory and inquiry in American politics* (pp. 60-88). Columbus, Ohio: Ohio State University Press.

Somanathan, E., Sterner, T., Sugiyama, T., Chimanikire, D., Dubash, N. K., Essandoh-Yeddu, J., . . . Zylicz, T. (2014). National and Sub-national Policies

and Institutions. In O. Edenhofer, R. Pichs-Madruga, Y. Sokona, E. Farahani, S. Kadner, K. Seyboth, . . . J. C. Minx (Eds.), *Climate change 2014: mitigation of climate change. Contribution of working group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (pp. 1141-1205). Cambridge, UK; New York, USA: Cambridge University Press.

Sorrell, S., Smith, A., Betz, R., Walz, R., Boemare, C., Quirion, P., . . . Vassos, S. (2003). Interaction in EU climate policy, final report to DG Research under the Framework V project interaction in EU climate policy. Retrieved from http://sro.sussex.ac.uk/53992/1/INTERACT_Final_Report.pdf

Sprague, J. (1982). Is there a micro theory consistent with contextual analysis? In E. Ostrom (Ed.), *Strategies of political inquiry* (pp. 99-121). Beverly Hills: Sage.

Stame, N. (2003). Evaluation and the policy context: The European experience. *Evaluation Journal of Australasia*, *3*(2), 37-43.

Stame, N. (2006). Governance, democracy and evaluation. *Evaluation*, *12*(1), 7-16.

Stame, N. (2008). The European project, federalism and evaluation. *Evaluation*, *14*(2), 117-140.

Stern, E. (2009). Evaluation policy in the European Union and its institutions. *New Directions for Evaluation, 2009*(123), 67-85.

Stewart, R. B., Oppenheimer, M., & Rudyk, B. (2013). A new strategy for global climate protection. *Climatic Change*, *120*(1-2), 1-12.

Stockmann, R. (2006). Evaluation in Deutschland. In R. Stockmann (Ed.), *Evaluationsforschung: Grundlagen und ausgewählte Forschungsfelder* (pp. 11-40). Münster; New York; München; Wien: Waxmann.

Stockmann, R., & Meyer, W. (2014). *Evaluation: Eine Einführung*. Opladen: Budrich.

Struhkamp, G. (2007). Evaluation in Germany: An overview. *Journal of Multicisciplinary Evaluation, 2*(3), 180-194.

Sullivan, H. (2011). 'Truth' junkies: Using evaluation in UK public policy. *Policy and Politics, 39*(4), 499-512.

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

Tarko, V. (2017). *Elinor Ostrom: An intellectual biography*. Lanham, USA: Rowman & Littlefield.

Taylor, S., Bachtler, J., & Polverari, L. (2001). Structural fund evaluation as a programme management tool: Comparative assessment and reflections on Germany. *Informationen Zur Raumentwicklung, 6*(7), 341-357.

Technopolis. (2008). Technopolis effectiveness evaluation of the European Environment Agency. Retrieved from http://www.eea.europa.eu/about-us/governance/eea-evaluations/2008

The Economist. (2014). The Environmental Union: On climate change, if little else, Europe still aspires to global leadership. Retrieved from https://www.economist.com/news/europe/21629387-climate-change-if-little-else-europe-still-aspires-global-leadership-environmental

Thompson, D. F. (2008). Deliberative democratic theory and empirical political science. *Annual Review of Political Science, 11*, 497-520.

Thompson, T. M., Rausch, S., Saari, R. K., & Selin, N. E. (2014). A systems approach to evaluating the air quality co-benefits of US carbon policies. *Nature*

Climate Change, *4*, 917-923. Retrieved from http://dx.doi.org/10.1038/nclimate2342

Tilly, C., & Goodin, R. E. (2006). It depends. In R. E. Goodin, & C. Tilly (Eds.), *The Oxford handbook of contextual political analysis* (pp. 3-32). Oxford; New York: Oxford University Press.

Tosun, J., & Schoenefeld, J. J. (2017). Collective climate action and networked climate governance. *Wiley Interdisciplinary Reviews: Climate Change*, *8*(1) Retrieved from https://doi.org/10.1002/wcc.440

Toulemonde, J. (2000). Evaluation culture (s) in Europe: Differences and convergence between national practices. *Vierteljahrshefte Zur Wirtschaftsforschung*, *69*(3), 350-357.

Turnpenny, J., Russel, D., Jordan, A., Bond, A., & Sheate, W. R. (2016). Environment. In C. A. Dunlop, & C. M. Radaelli (Eds.), *Handbook of regulatory impact assessment* (pp. 193-208). Cheltenham: Edward Elgar Publishing.

Uitto, J. I. (2016). Evaluating the environment as a global public good. *Evaluation*, *22*(1), 108-115.

UKES 2013 Colin Jacobs. UK Evaluation Society (Director). (2013).[Video/DVD] Retrieved from https://www.youtube.com/watch?v=17_aLZUQglI

van Voorst, S. (2017). Evaluation capacity in the European Commission. *Evaluation*, *23*(1), 24-41.

van Voorst, S., & Zwaan, P. (2018). The (non-) use of ex post legislative evaluations by the European Commission. *Journal of European Public Policy*, doi:DOI: 10.1080/13501763.2018.1449235

Vedung, E. (2013). Six models of evaluation. In E. Araral, S. Fritzen, M. Howlett, M. Ramesh & X. Wu (Eds.), *Routledge handbook of public policy* (pp. 387-400). London; New York: Routledge.

Vedung, E. (1997). *Public policy and program evaluation*. New Bruswick, N.J.: Transaction Publishers.

Versluis, E., van Keulen, M., & Stephenson, P. (2011). *Analyzing the European Union policy process*. Basingstoke: Palgrave Macmillan.

Victor, D. G., House, J. C., & Joy, S. (2005). A Madisonian approach to climate policy. *Science, 309*(5742), 1820-1821.

Vo, A. T., & Christie, C. A. (2015). Advancing research on evaluation through the study of context. *New Directions for Evaluation, 2015*(148), 43-55.

Warren, P. (2014). A review of demand-side management policy in the UK. *Renewable and Sustainable Energy Reviews, 29*, 941-951.

Weiss, C. H. (1993). Where politics and evaluation research meet. *Evaluation Practice, 14*(1), 93-106.

Wells, P. (2007). New Labour and evidence based policy making: 1997-2007. *People, Place & Policy Online, 1*(1), 22-29.

Widmer, T. (2004). The development and status of evaluation standards in Western Europe. *New Directions for Evaluation, 2004*(104), 31-42.

Wirths, D., Rosser, C., Horber-Papazian, K., & Mader, L. (2017). Über die gesetzliche Verankerung von Evaluation: Die Verteilung von Evaluationsklauseln und deren Auswirkungen auf kantonaler Ebene. In F. Sager, T. Widmer & A. Balthasar (Eds.), *Evaluation im politischen System der Schweiz* (pp. 155-188). Zürich: NZZ Libro.

Wörlen, C. (2011). Meta-evaluation of climate mitigation evaluations. Retrieved from https://www.climate-eval.org/sites/default/files/studies/Climate-Eval%20Meta-Evaluation%20of%20Climate%20Mitigation%20Evaluations.pdf

Wörlen, C., Rieseberg, S., & Lorenz, R. (2014). *A national experiment without evaluation or monitoring and evaluating the Energiewende?*. Berlin: International Energy Policy and Programme Evaluation Conference.

Wurzel, R. K. W., & Connelly, J. (2011). *The European Union as a leader in international climate change politics*. Oxon; New York: Routledge.

Wyns, T. (2015). Lessons from the EU's ETS for a new international climate agreement. *The International Spectator*, *50*(1), 46-59.

Yamin, F., & Depledge, J. (2004). *The international climate change regime: a guide to rules, institutions and procedures*. Cambridge, UK; New York: Cambridge University Press.

Zito, A. R. (2009). European agencies as agents of governance and EU learning. *Journal of European Public Policy*, *16*(8), 1224-1243.

Zito, A. R., & Schout, A. (2009). Learning theory reconsidered: EU integration theories and learning. *Journal of European Public Policy*, *16*(8), 1103-1123.

Zwaan, P., van Voorst, S., & Mastenbroek, E. (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.

Appendix 1

Figure A1.1: Use of the terms ‘polycentricity’, ‘polycentrism’ and ‘polycentricism’ in Google Books, 1940-2008[82]

Source: This graph was generated using the Google Books NGram technology. For further information, see http://books.google.com/ngrams and (Michel et al., 2011); y-axis represents % of all n-grams (in this case, one-word unigrams, or search terms).

[82] Data were available from 1800 – 2008 at the time of writing; given virtually no use of the search terms before 1940, I chose to display 1940-2008.

Appendix 2

Table A2.1: Climate policy evaluation database: list of sources

Organization type	Germany	United Kingdom	EU level
Government	• Federal Ministry for the Environment, Nature Conservation, Building and Nuclear Safety • Federal Ministry for Economic Affairs and Energy • Federal Ministry of Food and Agriculture • Inter-ministerial working group on CO2 reduction	• Department of Energy and Climate Change (DECC) • DEFRA • HMRC • Department for Transport	• European Commission o DG Clima o DG Regio o DG Energy o DG Trade o DG Move o DG Agriculture o DG Environment o DG Health & Food Safety o DG Mare o DG Growth (Internal Market) o Joint Research Centre (JRC) • European Council
Parliaments & Parliamentary Bodies	• German Bundestag • German Bundesrat	• House of Commons Environmental Audit Committee • House of Commons EFRA committee	• European Parliament o Committee on the Environment, Public

	• German Bundestag – Environment Committee • German Bundestag – Economy and Energy Committee • German Bundestag – Food and Agriculture Committee • German Bundestag – Transport & Digital Infrastructure Committee	• House of Lords Science and Technology Committee • House of Commons Science and Technology Committee • House of Commons Energy and Climate Change Committee • House of Commons Science and Technology Committee • House of Commons Transport Committee • House of Lords EU Select Committee and Energy and Environment Sub-Committee	Health and Food Safety ○ Committee on Industry, Research and Energy ○ Temporary Committee on Climate Change ○ Committee on Transport and Tourism ○ Committee on Agriculture and Rural Development • European Parliament Research Service
Government agency/bank	• Sachverständigenrat für Umweltfragen (German Advisory Council on the Environment) • National Climate Protection Initiative • German Environmental Protection Agency (UBA) • German Energy Agency • German Advisory Council on Economy and Energy • German Advisory Council on Global Change	• UK Environment Agency • Low Carbon Innovation Coordination Group • UK Sustainable Development Commission • Carbon Trust • Ofgem • Energy Savings Trust • UK Climate Change Committee • Renewable Fuels Agency • Green Investment Bank	• European Environment Agency

	• German Advisory Council for Agriculture, Food and Consumer Protection • KFW Bank		
Courts & Scrutiny Bodies	• Bundesrechnungshof	• UK National Audit Office • UK Audit Commission	• European Court of Auditors
Private consultancy	• Ecofys • Arepo Consult • Prognos AG • GfK	• Ipsos Mori • Eunomia • Frontier Economics • Ecofys • Golder Associates • Cambridge Econometrics • Future Energy Solutions • Enviros Consulting • Oxera Consulting • McKinsey • IPA Advisory Limited • NERA Economic Consulting • Ricardo Consulting • Ricardo Energy & Environment • LEK Consulting • Eoin Lees Energy • ADAS Consulting • Phylipsen Climate Change Consulting • ICF International	• Ecofys
Non-governmental organization	• Germanwatch • Öko Institute	• UK Green Alliance • Carbon Connect	• Open Europe • Greenpeace (EU)

	• ADAC Germany • Greenpeace (Germany) • WWF Germany • BUND (Friends of the Earth Germany)	• Centre for Sustainable Energy • Sandbag • WWF UK • Greenpeace UK • Friends of the Earth UK • E3G • Policy Exchange • What about those affiliated to political parties, eg SERA?	• Transport & Environment • European Environmental Bureau • Friends of the Earth Europe • Climate Action Network Europe • WWF European Policy Office • CEE Bankwatch Network
Private Sector Interest Group	• The Federation of German Industry (BDI) • Bundesverband Erneuerbare Energien e.V.	• The Confederation of British Industry • Engineering Employers Federation • Energy UK	• Business Europe • European Wind Energy Association • ePURE (European Renewable Ethanol) • AEBIOM (European Biomass Association) • EUREC (The Association of European Renewable Energy Centres) • ACEA (European Automobile Manufacturers Association) • Verband der Automobilindustrie (German Automobile Manufacturers Association)
Research institute/university	• Ecologic Institute • Fraunhofer Institute	• Institute of European Environmental Policy • UK Energy Research Centre	• Institute of European Environmental Policy (IEEP)

	• German Institute for Economic Research (DIW) • Bremen Energy Institute • Centre for Solar Energy and Hydrogen Research • Rhine-Westphalian Institute for Economic Research (RWI) • Forschungszentrum Jülich • Institute for Resource Efficiency and Energy Strategies (IREES) • Institute for Ecological Economic Research (IÖW) • Kiel Institute for the World Economy	• Grantham Institute • Tyndall Centre • UK Climate Impacts Programme (UKCIP) • NatCen • Policy Studies Institute • UCL Energy Institute • Environmental Change Institute • UCL Environment Institute • Centre for Sustainable Energy • Institute for Public Policy Research	• Centre for European Policy Studies (CEPS) • Centre for European Economic Research (ZEW)
Existing database	• ADAM Project database • Monitoring Mechanism report to European Environment Agency • Energiewende Studies Database (Forschungsradar)	• ADAM Project database • Monitoring Mechanism report to European Environment Agency • UK National Archives • Warren (2014) on demand-side response studies	• ADAM Project database • Commission Smart Regulation/Evaluation Database • Mastenbroek et al. (2015) database (publication in JEPP) • Commission Multi-Annual Overview of Evaluations & Impact Assessments (2002-2009) • Eureval Database • EU Climate Policy Biography (EUI)

Other	• Heinrich Böll Foundation • Konrad Adenauer Stiftung • Rosa Luxembourg Stiftung • Friedrich Ebert Stiftung • Friedrich Naumann Stiftung	• Anglo-German Foundation	• EUFORES (The European Forum for Renewable Energy Sources)

Appendix 3

Table A3.1: The Coding Scheme

Sub-dimensions - Potential operationalization	Coding: i.e. how will each sub-dimension be done in practice?	Score	Examples/page numbers for evidence.
Time (historical developments)	Length of time (retrospective) considered in the evaluation: 1 = (snapshot evaluation); 2 = >0 – 5 years; 3 = 6-10 years; 4 = 11-15 years; 5 = 16 – 20 years; 6 = >20 years		
Policy goals (intended outcomes)	Scale (separate evaluation for each dimension) – 5 point *Likert* with the following anchors: 0 = No reference to dimension (e.g. history, policy goals, etc.) 1 = Dimension discussed, but not explanation of how this dimension impacts policy outcomes (e.g., as part of general intro/overview) 2 = Dimension discussed, but limited explanation of how this dimension impacts policy outcomes (e.g., as part of general overview/intro) 3 = Dimension discussed, and good explanation of how this dimension impacts policy outcomes (e.g., as part of general overview/intro)		

	4 = Dimension discussed + impact on policy outcomes evaluated		
Policies in other sectors (interactions?)	See above.		
Unintended policy outcome(s)	See above.		
External events/circumstances	See above.		
Political environment/structures	See above.		
Geography	See above.		
Scientific findings (e.g., climate science)	See above.		
Evaluation methods used (list all) (Categorical)	Document analysis (lit review, case study) Modelling CBA Questionnaire/interviews Social experiment Stakeholder review/input Expert review/input Other/not specified.		
Number of methods used	Record number.		
Evaluation criteria used: (Categorical)	Effectiveness & goal attainment Coordination w/ other policies Fairness (incl. windfall profits, equity) Cost effectiveness Efficiency (incl. competitiveness, price impact) Legitimacy Accountability		

	Transparency Other/not specified		
Number of evaluation criteria.	Record number.		
Side effects	Yes/no		
Evaluation method: To what extent is the evaluation 'tailored' to the particular context in which the policy is placed? Routine instrument or fine-tuned?	Likert Scale 0 = No fine tuning (off-the-shelf method/model) 1 = Some fine tuning (e.g., general approach used, but some minor adjustments; e.g., some model calibration or choice of methods) 2 = Extensive fine-funding (more complex methods chosen/developed with a strong view to contextual effects/specific question)		
Reflexivity: existing policy targets considered as 'given' or critically evaluated?	Likert scale 1 = Policy goals accepted as given criteria for evaluation 2 = Policy goals critically questioned 3 = Policy goals critically questioned + alternative goals proposed		
Stated or implied purpose of the evaluation (Categorical)	1 = To satisfy a (legal) reporting requirement 2 = Accountability 3 = Improving policy/learning 4 = None/not clear		

Stated or implied target audience of the evaluation (Categorical; multiple mentions possible)	1 = Politicians (more general discussion) 2 = Policy-makers (bureaucrats) – fairly technical discussion 3 = The wider public 4 = Not clear		
Is the evaluation a legal requirement?	1 = Yes 0 = No		
Evaluation a continuous or one-off activity?	0 = No evidence of prior/subsequent evaluation activities 1 = Evidence of prior/subsequent evaluation activities		
For informal evaluation: Does informal evaluation attempt to identify and fill gaps left by 'formal' evaluation activities?	Likert scale 0 = No gaps in 'formal' evaluation identified 1 = Gaps identified 2 = Gaps identified and addressed		
Reference to evaluation studies conducted in other centres (but focusing on the same centre)	Likert scale 0 = No reference 1 = Limited, shallow reference (e.g., briefly mentioned) 2 = Good attention to studies from other goverance centres 3 = Extensive attention to other studies (description of the study/findings AND connection with analysis of the 'own' governance centre)		
Reference to evaluation studies focusing on or general experiences in other centres.	Likert scale 0 = No reference 1 = Limited, shallow reference (e.g., briefly mentioned)		

	2 = Good attention to studies focusing on other governance centres 3 = Extensive attention to other governance centres (description of the study/findings AND connection with analysis of the 'own' governance centre)		
To what extent do 'formal' evaluations draw on information from 'informal' evaluation and vice versa?	Likert scale 0 = no use of informal/formal data 1 = some use of formal/informal data (e.g., background) 2 = extensive use of formal/informal data (e.g., data used in core analyses)		
Is there some common metric (e.g., quantification?) that can be used to compare across governance centres?	Likert scale 0 = no common metric used 1 = 1 common metric used 2 = 2 common metrics used 3 = 3 common metrics used 4 = 4+ common metrics used		
Are there key lessons/Recommendations for others or the policy itself?	Likert scale 0 = No recommendations 1 = Some recommendations 2 = Extensive recommendations		
If there are recommendations, it is clear whether/how the context matters?	Likert scale 0 = No contextualization of recommendations 1 = Some contextualizations of recommendations 2 = Extensive contextualizations of recommendations		
Ease of use	0 = No executive summary 1 = Executive summary included		

If executive summary: (Categorical)	0 = no hierarchy of information (e.g., single block of text) 1 = hierarchy of information included (sub-headings, bolding, figures, etc.)		
Linguistic access	0 = No summary in other language 1 = Summary in other language included		
Availability of the evaluation	Likert scale 0 = Published without any efforts to publicize 1 = Some effort to publicize (e.g., available on website) 2 = Press release/policy brief 3 = Other governance centres actively informed about evaluation outcomes (e.g., press conference, available in multiple outlets, picked up by the media)?		
Evaluand – substance or process? (Categorical)	1 = policy substance 2 = policy process 3 = substance + process		
Evaluation budget	Record budget figure if available.		
General comments			

Appendix 4

Figure A4.1: Pearson correlations among the contextual variables (formal evaluations)

Variables	1	2	3	4	5	6	7	8
1. Time horizon	1							
2. Policy goals	0,023	1						
3. Other sectors	0.213	.256*	1					
4. Unintended policy outcomes	0.053	-0.074	0.057	1				
5. External events and circumstances	.255*	0.189	.237*	0.048	1			
6. Political environment	.301**	0.168	0.179	0.110	0.200	1		
7. Geography	0.173	0.001	0.156	.231*	0.034	.316**	1	
8. Science	.269*	-0.023	0.037	-0.022	0.113	.238*	0.208	1

*. Correlation is significant at the 0.05 level (2-tailed).

**. Correlation is significant at the 0.01 level (2-tailed).

Figure A4.2: Pearson correlations among the contextual variables (informal evaluations)

Variables	1	2	3	4	5	6	7	8
1. Time horizon	1							
2. Policy goals	0.091	1						
3. Other sectors	-0.051	-0.019	1					
4. Unintended policy outcomes	-0.002	-,331**	-0.006	1				
5. External events and circumstances	.282**	-0.072	0.166	0.038	1			
6. Political environment	0.152	,270*	,219*	-,323**	-0,051	1		
7. Geography	0,106	0,058	-0,054	0,052	0,108	0,160	1	
8. Science	0,185	0,101	0,174	0,064	0,011	-0,066	-0,109	1

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).