# 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." Accordingly, "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, pp. 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 actors can and do self-organize in order to address pressing governance challenges, that context matters in governance, and that governance centers, 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 analyze what role evaluation may play to contribute to the shift from polycentricity to polycentrism with a view to the foundational ideas. The chapter ultimately combines polycentric governance and policy evaluation literatures in novel ways to advance polycentric governance and policy evaluation theory. It concludes by further elaborating and specifying the key empirical research gaps, which Chapter 1 already flagged, and which will be addressed later.



https://doi.org/10.1017/9781009049658.003 Published online by Cambridge University Press

## 2.2 Positivism, Key Variables, and Normative Theory

From the outset, and as Chapter 1 recognized, it is critical to appreciate that theory on polycentrism contains a subtle blend of both normative and positive elements (Aligica, 2014; McGinnis, 2016). In his early and later republished 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 (Ostrom, 1999a). In more recent contributions, the normative element has become even more pronounced, as for example Michael D. McGinnis (2016) 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*[1] 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"[2] for common pool resource governance systems have since found further empirical support around the world (Ostrom, 2005). Given decades of empirical work and particularly Elinor Ostrom's affinity for "grounded research" and interdisciplinarity (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 the approach 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. Therefore, "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

[1] https://ostromworkshop.indiana.edu/.

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

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

arrangements" (Aligica, 2014, p. 17). This book thus endeavors 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 Elinor and Vincent Ostrom's general scholarly approach, which seeks to elevate theory over methods (Aligica & Sabetti, 2014b, p. 2). This approach reacts to the positivist doctrine starting in the 1960s, where scholars endeavored to build theory starting from empirical insights (for a treatment of this intellectual history, see Ostrom, 2014a). Therefore, 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 Ostrom also had a strong affinity for empirical work. She argues that while "theory precedes empirical work . . ., empirical studies help to refine our theoretical understanding of the world" (Ostrom, 2014a, p. 222). In sum, 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 Section 2.1), 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 (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 (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. If these conditions are met and actors self-organize, outcomes may be normatively "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 behavior – the key idea of subsidiarity. 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 or property rights (Hardin, 1968) does not withstand empirical scrutiny across all cases (Ostrom, 1990), although Elinor Ostrom recognizes potential drawbacks of polycentric governance arrangements, such as the possibility of free riding and potential under-provision of public goods (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 (Ostrom, 2010c). Thus, in some cases, actors appear to exhibit the capacity to self-organize and outperform top-down solutions. This book 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

Michael D. 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 (Ostrom et al., 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 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" (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, 2014a).

The third foundational idea holds that if polycentrism is to emerge, governance centers need to interact, but without generating strong interdependencies. But what is a "governance center"? 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 et al. (1961, p. 831) write about "centers 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 (Ostrom, 2005, p. 269), Elinor Ostrom 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 center" remains unclear, as it may range from individuals to all types of organizations or enterprises all the way to more fuzzily described "centers of decision-making." To complicate things more, a recurring theme in Elinor Ostrom's scholarship is that governance centers are "nested" (see quote above), which creates the challenge to not only tell governance centers 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 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 center. This book defines governance center 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 Ostrom et al., 1961). This definition therefore encompasses the level of the nation state and supranational organizations like the EU.

Linked to the idea of "nesting," what drives interactions between centers of governance in polycentric systems? There are numerous potential mechanisms. Vincent Ostrom believed that governance centers will interact more or less automatically if they have sufficient incentives to do so (Ostrom et al., 1961). Overlapping jurisdictions may be one reason why centers interact. For example, writing on the IAD, Michael D. 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 rulemaking venues in a polycentric system. These individuals or organizations become bridges between different governance centers to foster interaction. In other cases, interaction may emerge because of market-like competition – for example, when different governance centers 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 favorable, depending on the dimension that matters most to them (e.g., cost; quality of the service). 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 kindle interaction. By definition, the polycentric approach excludes ideas around top-down coercion, as governance centers are per definition thought to be independent.

In climate change governance, the threat of "carbon leakage" provides one potential (external and market-driven) incentive for governance centers 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 rules in a classic "race to the bottom" (Ostrom, 2014b). Thus, if public policymakers 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 centers take equivalent action.

An additional reason to look beyond one's own governance center is to learn from the successes and failures of others, especially because policymakers 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 behavior or beliefs, following the impact of experience, new information, or changing circumstances (Bennett & Howlett, 1992). Of particular interest to this book 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 socially desirable side effects, such as improvements in human health or reducing congestion (Thompson et al., 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 centers. 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 policymakers 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 centers may be especially interested in the experiments of others as a key source of lessons (Aligica, 2014, p. 66; Goodin, 1996; Hildén et al., 2017).

An issue, of course, emerges with regard to the previous points about context. If context matters in policymaking, 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 centers. 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. 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 tends to 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 center 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 voluminous 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 centers 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 behavior, preexisting institutions, and general political inertia bolstered by overburdened policymakers, 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 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), the key issue that polycentric governance seeks to address (Ostrom, 1990). Even though the system as a whole could benefit from learning, individual governance centers may not be able or willing to draw lessons from others or to provide their own lessons. For these reasons, some higher-level incentives, if only through coordination, may be necessary to drive interaction in

polycentric systems (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" (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 preexisting 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" (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 preexisting 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 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 goes on to argue that more traditional public actors including states may deliver some or all of these four functions. Notably, although Mansbridge 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 requires conscious effort and coordination. These questions have a direct bearing on the central questions of this book, 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 centers (and thus the shift from polycentricity to polycentrism that Chapter 1 explains).

Crucially for this book, 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 (2014) 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 centers 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 centers on who provides "collective" or "public" goods, which may include the extent to which governance centers 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

https://doi.org/10.1017/9781009049658.003 Published online by Cambridge University Press

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, pp. 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 Section 2.3). 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 (Ostrom & Nagendra, 2007). Participants may then even be willing to bear some of the cost of monitoring themselves (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 centers – 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*" (Ostrom, 1990, 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" (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 (Leeuw, 2010; Ostrom, 1990; Schoenefeld et al., 2019, 2021; Schoenefeld & Jordan, 2020). 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 (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" (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" (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" (Ostrom, 1990, p. 203). And if frequent monitoring is required, costs tend to increase (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 and why" (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 costlier to enforce" (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. *(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" (Ostrom, 2005, p. 283). Self-organization may in turn support interactions between governance centers.

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 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 centers.

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

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.<br>• 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 resources are much more difficult to monitor than others.<br>• Larger systems are more difficult to monitor than smaller ones.<br>• Precise rules are easier to monitor than more general ones.<br>• 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 centers; this can lead to learning from different experiments. |

*Sources:* Ostrom (1990), Ostrom (2005), Ostrom and Nagendra (2007).

of monitoring decisions. For example, Canada and the EU have quarreled intensely about the greenhouse gas content of tar sand oil (e.g., 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 et al., 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, the physical nature of greenhouse gases makes monitoring their emissions 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 center of policy discussions – and the "goals" of a policy may indeed be subject to significant contestation. Similarly, supply-oriented climate policy aims to leave significant amounts 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). 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. 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; Schoenefeld et al., 2019) that top-down

monitoring has proven difficult if not impossible to install 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 behavior of her or his colleagues only has limited value when considering the monitoring of national, regional, or even local 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. The same is also true for some subnational actors, as monitoring at the city or the company level has shown. 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 (Mor & Ghimire, 2022; 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 centers.

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 centers is part and parcel of polycentric governance. Empirical research on monitoring in common-pool resource systems (see above Section 2.4) 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

https://doi.org/10.1017/9781009049658.003 Published online by Cambridge University Press

Table 2.2 *Monitoring (public) climate policy*

| | |
|---|---|
| Self-organization | • Self-monitoring can happen at national and subnational levels (by both state and nonstate actors).<br>• 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.<br>• The "climate system" is very large (global).<br>• It is difficult to define clear-cut and precise rules for monitoring, given technical issues and political sensitivities.<br>• 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 centers; this can lead to learning from different experiments; this can also happen at the international level, e.g. EU – see Schoenefeld et al. (2018). |

substantially different activity. Recall that this book 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 detect.

Policy evaluation is thus a broader activity than monitoring, and therefore, its role in polycentric governance system must also be considered in broader terms (see Schoenefeld & Jordan, 2019). 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 et al., 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

https://doi.org/10.1017/9781009049658.003 Published online by Cambridge University Press

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 policymakers 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 play a role 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 book 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 centers (see above Section 2.3).

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; Lascoumes & Le Gales, 2007; Nilsson et al., 2008; Owens et al., 2004; Vedung, 1997). First, Weiss (1993) argued that because government programs 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 in Section 2.3). For example, a sympathetic 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 (Gravey & Jordan, 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 (see Schoenefeld & Jordan, 2019).

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 its outcome 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 Section 2.5.1). For example, society-driven actors may conduct or commission their "own" evaluations in order to contest points made by state-driven 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 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 centers. 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, state and societal 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 *state-driven* and *society-driven* 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 policymaking 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 EU (Hildén et al., 2014; Huitema et al., 2011; Schoenefeld, 2021). Hildén et al. (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 who or what initiates evaluation. For example, Sager et al. (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 potential self-organizing capacities in scientific assessment, monitoring, and policy evaluation. 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 within the same organization. However, it should be noted that a footnote in the paper explains that many of the noncommissioned evaluations may have emerged from academic research projects, as this particular study used a wider operational definition of policy evaluation than the one applied in this book (see Chapter 1). 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 may add an extra dose of independence to evaluation (see Chelimsky, 2009), emerging research suggests that in practice, it can be a site of political struggle where those who commission evaluations often try to control their contractors (Pleger & Sager, 2018). 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 – for example, not enough funding for a control group (see Pleger & Sager, 2018 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, 2018), 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 state-driven and society-driven evaluation becomes significantly more difficult once multiple actors become involved in a single evaluation.

While evaluation literatures have long problematized the relationship between state-driven and society-driven evaluation and their influence on evaluation results, early scholars often considered state-driven and society-driven categories rather crudely, that is, paying insufficient attention to principal-agent relationships between evaluation funders and evaluators whenever evaluations are commissioned (Weiss, 1993). Newer evidence challenges this 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 were 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 state and societal evaluators commission evaluations. For example, Hayward et al. (2013) consider this principal-agent relationship and show how (state-driven) 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 societal actors have a particularly critical role to play in evaluation 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 [nongovernmental organizations] 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 societal 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 empower themselves (Fetterman & Wandersman, 2005). The approach has devout followers – for example, Diaz-Puente et al. (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). Not all participation is equally "empowering." For example, individuals may simply be asked how satisfied they are with a particular service, generating more passive involvement. Other approaches in the participatory tradition go further 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 legal requirements to evaluate, often in the form of "evaluation clauses" 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 (see also Schoenefeld, 2021). For example, Mastenbroek et al. (2016) found that out of the 216 European Commission ex post legislative evaluations they identified, 81 percent 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 ninety such clauses at the national level in Switzerland (Bussmann, 2008). Evaluation scholars have taken the increasing presence of evaluation clauses as an indication of advanced evaluation institutionalization (Jacob et al., 2015). This book 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 state/society-driven distinction has proven a useful conceptual tool, it remains an open question to what extent the categories of state and society-based evaluators blur or even interact, as has been suggested in other policy areas (Guha-Khasnobis et al., 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 endeavor. The following section focuses on this core issue.

### 2.5.2 Context

The idea that context matters in policy evaluation is not new, but 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 is 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) 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" (p. 213). 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 endeavors (Pawson & Tilley, 1997, pp. 63–78). More fundamentally, Guba and Lincoln (1981, pp. 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 sulfur dioxide trading in the USA 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 et al., 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, p. 27) 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." 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 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 policy development pathways, 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 Schoenefeld & Jordan, 2019; and Zito & Schout, 2009 for a discussion of different types of learning). Given the clear arguments on context by polycentric governance scholars this book 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 policy evaluations may account for context is through the evaluation approach, for example through evaluation methodologies or criteria. The second part of this section thus turns to the relevant discussions in this area.

With a view to 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) proposed 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, 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 et al., 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 book 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 policy.

***Time:*** While there may have been a time when scholars considered policymaking largely atemporal and independent of the effects of time, more recent discussions in public policy and management have sought to reintroduce the variable of time (see Pollitt, 2008). These general debates have also been addressed in the context of policy evaluation. For example, Bressers et al. (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,[3] 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, 2003, 2013). 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 counterproductive. Considering 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 considering 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 policymaking process (pp. 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: 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

[3] These numbers may look different when considering full life-cycle emissions.

https://doi.org/10.1017/9781009049658.003 Published online by Cambridge University Press

extent depend on the characteristics of the governance center 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 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 runs 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 analyze these dimensions. Thus, paying close attention to the physical, but also the sociopolitical 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 functions in unison with others (such as spatial planning policies to accompany 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, policymakers 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, 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, pp. 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, pp. 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, pp. 241–245) explains, policy outcomes also likely depend on interactions with other policies (sometimes strengthening, sometimes detracting from the policy), 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 discussion earlier in this section), 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 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: Frank 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" (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, Frank 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; Schoenefeld, 2023). 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.* *(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 list should thus be understood as a starting point for the empirical investigation (see the following chapters in this book), 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 light on important co-benefits at varying scales in addition to reducing global carbon dioxide emissions (Ostrom, 2010b; 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 centers take each other into account and, ideally, learn from each other. Learning from each other necessitates some mechanism through which governance centers 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 (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 centers 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 centers. Various factors may make lesson-learning across governance centers more or less likely. A crucial first step is that policy evaluations must become available to other governance centers in order to be able to have an effect. When governance actors can easily obtain evaluations from other governance centers (e.g., 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) policymakers to quickly assess whether an evaluation may be relevant to their situation (Zwaan et al., 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 center 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 centers matters in this respect.

And yet reverting back to the debate on idiosyncratic evaluation criteria and generalization raises key and difficult questions about comparability and thus learning opportunities (Schoenefeld et al., 2018; Schoenefeld & Jordan, 2017). A combination of providing both contextual analysis that considers contextual effects, but also some more general criteria or metrics that enable comparison seems of order. Aligica and Sabetti (2014a) 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 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 centers, 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 centers is relatively new and has surprisingly been little discussed in evaluation literatures.

Then there is the potential interaction between state and society-driven evaluation activities. As discussed earlier, scholars have developed the distinction between state and societal actors in evaluation. But how do the state and the societal 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 (Christiansen & Neuhold, 2013; Kleine, 2013). These literatures suggest that the interaction between state and society 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, state 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 sulfur 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 about learning, policy evaluation may also help governance centers 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 centers to contribute to accountability mechanisms. In addition, knowledge flowing from evaluation may, to a certain extent, also enable competition between governance centers (see 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 may also make evaluation actors reluctant to publicize their findings, particularly when they describe key factors that drive success (or, potentially, failure).

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 and related disciplines (e.g., Albaek, 1995; Haas, 2004; Radaelli, 1995; Rich, 1997) for space and practical reasons this book considers the more circumscribed *intended* target audiences (and thus potential users) of an evaluation.[4] Intended evaluation users are often policymakers, 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 restricted circles of individuals. 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 conducted by policymakers 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 centers 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 Section 2.5). Another function is to provide some accountability in governance settings where traditional accountability chains have been weakened or no longer exist (see Bäckstrand et al., 2018). In this understanding, evaluation becomes in effect a public good, which is

[4] Note that the extensive debates of (evaluation) knowledge utilization (e.g., Johnson et al., 2009; Rich, 1991) thus remain, by and large, outside the scope of this book.

nonexclusionary (if evaluations are public) and nonrivalrous (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 endeavored 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 to this question, yet often ill developed and with little 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 centers. But to date there is a severe paucity of studies that consider these factors together. 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 Elinor and Vincent 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. The foundational insights are: (1) context matters in governance; (2) actors can and sometimes do self-organize to muster governance solutions; and (3) interaction between otherwise independent governance centers 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.