Introduction: putting policy evaluation into its democratic context

Frédéric Varone, Steve Jacob and Pirmin Bundi

THE INHERENT POLITICS OF POLICY EVALUATION

Policy evaluation is frequently depicted as the final stage in the heuristic model of the policy cycle (e.g., Hill & Varone, 2021, pp. 301ff.; Howlett et al., 2009). Once a public policy has been put on the agenda, designed, enacted and implemented, a policy evaluation aims at measuring the effects generated by this policy to ensure accountability in the policymaking process and to identify whether the policy has achieved its stated objectives. Elected politicians and civil servants can then adjust the policy design and implementation based on the evidence provided by the evaluation.

To realize this first ambition of a policy evaluation, that is, isolating policy effects, evaluators must conceptually and empirically reconstruct the complex causal mechanism from the policy's inception through to the implementation activities of public administration (outputs) and then on to the behavioural changes of target groups (outcomes) and, eventually, the desired effects of the policy on society (impacts). To meet this challenge, policy evaluators must have strong analytical and methodological skills.

Assuming that evaluators are able to define and measure policy effects accurately, the second ambition of a policy evaluation is to enable a value judgement of these policy outcomes and impacts. Evaluators must eventually judge the merit, worth, value or utility of a public policy (Scriven, 1991) with the aid of explicit evaluation criteria. The most common criteria used in policy evaluations focus on effectiveness (i.e., achievement of policy goals), efficiency (i.e., ratio between policy effects and resources, such as personnel, money and time) and economy (i.e., optimal use of resources to deliver policy outputs). Less frequently, but no less importantly, policy evaluations also question the relevance of a policy and examine the adequacy of the stated policy objectives to solve the social problem at hand and meet the needs of the policy beneficiaries. Many additional evaluation criteria, which will not be listed here, are also used in specific policy domains or institutional contexts (e.g., 'coherence' to assess the fit of a policy measure or 'sustainability' to evaluate whether the benefits will last; as suggested by the Organisation for Economic Co-operation and Development [OECD] DAC Network on Development Evaluation, n.d.).

The heroic postulate that a policymaking process should involve the explicit identification of policy objectives and the translation of these objectives into expected policy effects that can be empirically measured dies hard. The systematic search for and use of evaluation findings about 'what works, under what circumstances, and why' is laudable from a conceptual point of view. However, as attractive as it might be for evaluators and decision-makers, policymaking based on evidence from evaluation is not easy to implement in reality. Indeed, policy evaluations are only selectively conducted in practice; furthermore, they mostly have a limited influence on the policymaking process (see Christie, 2007).

The limited development and fragmentary usage of policy evaluations are partially rooted in epistemology and social science methods. For the staunch advocates of a positivist evaluation approach, an important problem is the obvious difficulty in developing a pure experimental or quasi-experimental situation in the real world to identify the causal or net effects of a policy – the first ambition of a policy evaluation (see Chapter 22 by John and Chapter 23 by Andersen). In addition to these methodological issues, one also needs to consider the political dimension that is inherent to any policy evaluation process and to judgement about the worth of policy effects – the second ambition of a policy evaluation (see Chapter 1 by Patton, and Chapter 8 by Fischer). Social constructivist scholars claim that an evaluation is always subjective, since evaluation does not study policy effects as such but rather their interpretation by policy stakeholders. The evaluation process itself is based on values and beliefs and is therefore socially constructed (Fischer, 1995/2005).

In other words, policymaking is not only a technical exercise consisting of selecting policy instruments according to their (expected) effects and just doing 'what works'. It inevitably involves trade-offs between multiple competing social values and related policy objectives. If policymaking is about 'who gets what, when, and how', as highlighted by Harold L. Lasswell in 1936, then evidence should primarily serve to capture who benefits from different policy choices and who does not (see Chapter 9 by Rey and Fortin, Chapter 10 by Pires and Lotta, and Chapter 11 by Mertens). However, evaluators cannot indicate which is the right policy choice, that is, which citizens and social groups deserve policy benefits and which do not. This is obviously a political choice that should be legitimated by elected decision-makers or by citizens if they are called to vote on public policy (see Chapter 16 by Sager, Schlaufer and Stucki). Empirical evidence about policy effects alone has no bearing on the social desirability, political acceptability and democratic legitimacy of what has been (accurately) measured (through experimental evaluation methods).

To overcome the confrontation between realists and constructivists (see Chapter 2 by Tosun, De Francesco and Pattyn, and Chapter 3 by Fontaine), innovative evaluation approaches have been proposed. For instance, Frank Fischer (1995/2005) suggests combining four dimensions in an 'empowering' evaluation process: (1) measuring the achievement of stated policy objectives (i.e., verification); (2) identifying issues about the relevance of the policy for the social problem it claims to address (i.e., validation); (3) asking whether the policy contributes value for society as a whole (i.e., vindication); and, finally (4) raising wider ideological questions about what the policy is trying to accomplish (i.e., social choice).

Our introduction to this *Handbook* will not further explore the challenges presented by Fischer's approach to policy evaluation, but rather highlight that evaluation is always a political activity (Eliadis et al., 2011; Weiss, 1975). Evaluators should be aware of the political forces shaping the evaluation scope and process, or their findings bear the risk of being instrumentalized to reinforce existing power relations (Bundi & Trein, 2022). This *Handbook* thus focuses on the political dimension of the policy evaluation process rather than the technical and methodological problems about how best to capture the net effects of a public policy. The aim is to put evaluation into its context within the policymaking and democratic process. The 26 *Handbook* chapters are divided into four parts, summarized below.

PART I: EVALUATION, ACCOUNTABILITY AND LEARNING IN THE POLICY PROCESS

The first part presents the role of policy evaluation in the policy process. It investigates how evaluation relates to policy design, law-making, budgetary processes, policy learning, and evidence-based policymaking. This part also compares policy evaluation to other instruments (e.g., performance auditing) used to influence and steer the policy process and service delivery by public administrations.

In Chapter 1, Michael Quinn Patton introduces the sociohistorical roots of policy evaluation and retraces its evolution, showing how modern policy evaluation has ancient precedents. Indeed, policy evaluation depends on and is rooted in rationality, the kind of critical thinking exemplified by the teachings of Socrates in ancient Greece, and his method of questioning in particular. Historical examples reveal the evolution of key elements of policy evaluation. The modern vision of an ‘experimenting society’ (Campbell, 1971/1991) in which policies are tested, improved and adapted through evaluation depends on its findings being used. Major barriers to evaluation use (House, 1972; Weiss, 1972; Wholey et al., 1970) have emerged, as have practices of misusing and distorting evaluation findings to serve political purposes, such as neoliberal retrenchment policies and new public management reforms (see also Chapter 6 by Peters and Pierre). Finally, Patton looks forward and discusses the evaluations of governmental responses to the COVID-19 pandemic. He suggests that policy evaluation is facing new challenges, since it must counter the anti-science trends of the post-truth era characterized by an ‘infodemic’ of misinformation, fake news and conspiracy theories (see also Chapter 8 by Fischer, and Chapter 22 by John). Hopefully, public policy evaluators will continue to create a balanced, informative and useful synthesis from disparate and often conflicting findings.

Jale Tosun, Fabrizio De Francesco and Valérie Pattyn also stress that policy evaluation is a political activity by nature. Although evaluators are embedded in political contexts, they must first develop a toolbox and routines that enable them to carry out policy evaluations. Consequently, the second chapter concentrates on two prominent evaluation approaches: the positivist and the social constructivist. Chapter 2 is guided by the following questions: How much social constructivism (Guba & Lincoln, 1989) exists in realist evaluation theory (Pawson & Tilley, 1997; see Chapter 3 by Fontaine)? How much in practice? Can both approaches be integrated without reducing their ontological, epistemological and methodological integrity? The authors eventually show that realist evaluation is conceptually open to social constructivism: ‘After all, realist evaluation can be seen as an approach to public policy and policy evaluation that is rooted in scientific realism (which contends that aspects of the world can be described by the sciences), but which at the same time is aware that human behaviour is not only the outcome of incentives provided by policy measures, but is also affected by context-specific factors of different kinds’.

Chapter 3, by Guillaume Fontaine, reviews the core concepts and methodological grounds of realist evaluation to operationalize the models of evaluation within the framework of policy design. After justifying the relevance of realist evaluation to public policy (Pawson & Tilley, 1997), the author explains how realist evaluation and policy design meet, particularly to address issues of context and causality. He presents the cornerstones of realist evaluation and key concepts such as theory-driven ‘interventions’, ‘systematic reviews’, ‘generative causation’, and ‘causal mechanism outcome configurations’. He then describes the main prospects and challenges of realist evaluation, focusing on the use of multi-methods for scaling down (for internal validity) and scaling up (for external validity). The chapter concludes with some reflections about the relationships between realist evaluation and policy design in both research and practice. Policy evaluation should be understood as a learning process that informs policy redesign (see also Chapter 4 by Flückiger and Popelier, and Chapter 5 by Dunlop and Radaelli).

In Chapter 4, Alexandre Flückiger and Patricia Popelier also claim that policy evaluation is better understood as a recurring activity than as the final stage of the policy cycle. Integrating both ex ante and ex post evaluations in the legislative process leads to better law-making. It allows legislators to focus on the actual impact of their legal and regulatory texts and to adapt them over time in response to their effects in a continuous learning and improving process. Legislations can thus produce optimal impact based on evidence, contributing to solving societal problems in a relevant, efficient, reflexive and fair way. Furthermore, the two legal scholars suggest that the courts should verify that this is the case in reality, which they accomplish mainly through the 'proportionality test'. Ideally, policy evaluations also help the courts to uphold fundamental rights and principles (see also Chapter 15 by St-Georges and Rothmayr Allison). Flückiger and Popelier note that in practice, however, the quality of regulatory impact assessments has often been reported as flawed, and legislators frequently do not take evaluation results seriously (see also Chapter 13 by Bundi). Nevertheless, bringing policy evaluation into the legislative process has already enabled the development of innovative types of legislation, such as temporary laws subject to review (i.e., sunset and experimental legislation), self-regulatory triggering laws (i.e., Damocles laws), and legislation by objectives (i.e., programmatic laws). In the future, artificial intelligence (AI) tools applied to legislation could possibly reshape the role of policy evaluation more fundamentally (see also Chapter 24 by Cahlikova and Ballester). Real-time monitoring of law implementation and, furthermore, automatic revision of the legal rules according to new facts communicated directly from the real world by appropriate algorithms would allow for the evaluation of policy effectiveness on an ongoing basis. Such a potential development would be highly problematic, since policy evaluation and the subsequent revision of legal rules require more than is possible to achieve with automatic decision-making.

If automated decisions are indeed the opposite of political choices, then one should focus on how evaluation findings can trigger learning mechanisms in the policy process and understand how policymakers update their beliefs and preferences about alternative policy options. As highlighted by Claire A. Dunlop and Claudio M. Radaelli, since its origin, policy evaluation has been geared towards the objective of policy learning. Chapter 5 reviews foundational approaches to policy evaluation and their causal mechanisms that lead to learning. They eventually distinguish between four ideal-typical evaluation contexts based on the policy actors driving the process of learning: (1) the epistemic context, in which evaluators – as authoritative experts – are commissioned to teach technical knowledge (i.e., 'instrumental evaluation use' according to Weiss, 1972); (2) the reflexive context, in which dialogue occurs via the active participation of citizens during policy evaluations (see Guba & Lincoln, 1989); (3) the bargaining context, in which evaluators foster exchange through the consultation of vested interest; and (4) the hierarchical context, in which policy is scrutinized via monitoring and sanctioning of poor performance.

The relationship between public performance and policy evaluation is at the core of Chapter 6 written by B. Guy Peters and Jon Pierre. The essential role of policy evaluation has been challenged over the past couple of decades of growing attention given to auditing (see Barrados & Lonsdale, 2021; Power, 1997). According to the authors, this development can be attributed both to an increasing focus on the performance of government bodies and to an expansion of the scope of auditing conducted by Supreme Auditing Institutions (SAIs). Evaluation as a professional and scholarly field has developed theories and advanced methods to assess the effectiveness of public programmes (see also Chapter 21 by Gauthier and Roy). The growth of auditing may thus change the focus and quality of policy evaluation. Drawing on observations from a number of advanced democracies, Peters and Pierre demonstrate how conventional auditing institutions have become increasingly concerned with assisting policy change and administrative reform in the public sector – tasks that were traditionally associated with policy evaluation (see also Chapter 12 by Jacob). At the same time, auditing has in many ways crowded out evaluation as an integral part of the policy process. The (potential) consequences of this development for the audited and reforming institutions as well as for policymaking are important. The growth of performance auditing in areas previously assigned to policy evaluators can lead to a shortened time perspective, a stronger emphasis on the administration of policies, and an increased focus on efficiency and economy of the audited entity, among other things. At the same time, the shift from in-house policy evaluation to performance assessment by independent auditing organizations has often made audit findings more open to the legislature and the public.

The last chapter of the first part deals with the contribution of policy evaluation to accountability mechanisms within a policy process. Bovens (2007, p. 450) 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 judgement, and the actor may face consequences’. Applying this generic definition to policy evaluation, Yousueng Han investigates how diverse stakeholders participate in policy evaluation and, thereby, contribute to the three steps of the accountability process – namely, providing performance information via policy evaluation, debating about this information, and either punishing or rewarding policy actors as a consequence. Chapter 7 discusses how these accountability mechanisms are relevant for individual bureaucrats (see also Chapter 25 by Thomann and Lieberherr) and public agencies, experts, elected politicians, and judges (see also Chapter 15 by St-Georges and Rothmayr Allison, and Chapter 13 by Bundi). Finally, it indicates that methodological flaws during policy evaluation lead to inaccurate estimation of policy effects and, consequently, jeopardize the contribution of policy evaluation to democratic accountability mechanisms.

PART II: EVALUATION, PUBLIC PARTICIPATION AND DEMOCRACY

The second part of the *Handbook* discusses the openness of the evaluation process and questions how policy evaluation may foster public participation, strengthen governance, and, eventually, enhance democracy. Special focus is placed on the contributions that (post-positivist) evaluation approaches may offer to accompany recent transformations of advanced democracies, such as the increase of social inequalities. The main question here is to what extent does policy evaluation contribute to managing key challenges of democratic regimes?

In Chapter 8, Frank Fischer stresses that policy evaluation raises issues pertaining to both empirical and normative analyses. Insofar as evaluation is designed to inform the real world of policy decision-making, it is important to find a way to integrate the outcomes of both types of analysis. Fischer's argumentative framework incorporates empirical and normative evaluation in a systematic approach that facilitates a discursive-analytic probe (Fischer, 1995/2005). Based on the logic of practical reason and founded from an interpretive perspective, this evaluation framework is conducive to the everyday mode of argumentation that organizes and mediates deliberative communication among elected politicians, policy implementers and citizens. The logic of practical reason is scarcely new – as already underlined by Patton in Chapter 1 – but it has largely been ignored by social science and (positivist) policy evaluators. The 'argumentative turn' in policy analysis (Fischer & Forester, 1993) and policy evaluation thus aims to better connect theory to practice and action. Fischer illustrates the use of such an evaluation framework with the contemporary case of COVID-19 policy.

Lynda Rey and Alexandre Fortin acknowledge that the new generation of argumentative evaluation schemes, starting with Guba and Lincoln (1989), represents a fundamental shift, which brings the importance of contextualization to the forefront. The authors invite a concerted, negotiated and co-constructed approach to the evaluation process. Chapter 9 illustrates how various approaches inspired by constructivist epistemology have thus emerged under the umbrella of participatory approaches to evaluation, focusing on the diversity of stakeholder perspectives and redefining the role of evaluators. However, Rey and Fortin also claim that the contribution of various types of participatory evaluation to the democratic process and outcomes is not obvious. After analysing the relevance and promises of stakeholder participation in the evaluation of public policies, they discuss the challenges of democratic deliberative evaluations, such as the selection for and depth of participation. The authors conclude by highlighting the complex journey for policy evaluators in considering social justice and decolonization approaches beyond participation and democracy. Evaluators must play the role of mediator and counsellor to facilitate the dialogue about policy effects, support collective learning, and foster the inclusion of marginalized groups.

The effects of policy evaluation on disadvantaged groups is precisely the topic of Chapter 10. Roberto Pires and Gabriela Lotta argue that when policies are put in practice, they often produce a series of effects other than those originally intended. Furthermore, some of these unintended effects may reinforce existing social inequalities, even when policies are formally aimed at alleviating them. In contexts in which social inequalities have been on the rise or remained largely stable, policy evaluation must play an important role in understanding the processes through which policy execution perpetuates inequalities (see also Chapter 26 by Daigneault). The authors address this topic from two complementary perspectives. First, they analyse how policy evaluation may become a risk for inequality reproduction when it is blind to unintended consequences of the policy process, further legitimizing policies that generate unrecognized forms of exclusion. Second, they propose an analytical tool that makes policy evaluation more attentive to the reproduction of social inequalities. Based on the empirical analysis of a large set of cases in which social inequalities were reproduced in policy processes, Pires and Lotta identify five dimensions in a practical roadmap to make policy evaluators more sensitive to the repercussions of policy (evaluation) processes to disadvantaged publics and to mitigate undesirable unintended effects. They also develop a research agenda at the interface of policy evaluation and the reproduction of social inequalities.

Donna M. Mertens also claims that public policies can sustain an oppressive status quo, as evidenced globally in the form of policies related to such areas as immigration, the economy, the environment, housing, health, safety and education. Increasing inequities are undermining democratic societies and leading to a lower quality of life for marginalized and vulnerable populations. Her key argument, which resonates with the content of the previous chapter by Pires and Lotta, is that 'being conscious about addressing inequities, power relationships, contextual factors, and uncovering oppressive realities is a necessary step to support the transformative change that is needed'. The use of a transformative lens for policy evaluation provides an avenue to develop or revise policies to create a more just world for oppressed and excluded social groups, including women; racial, ethnic, and religious groups; people living with disabilities; and LGBTQIA persons. Chapter 11 outlines six phases of a policy evaluation, applying transformative and mixed methods. It explores the positive impact of these methods using examples from across the globe.

PART III: INSTITUTIONALIZATION, PRACTICE AND PROFESSIONALIZATION OF POLICY EVALUATION

The third part of the *Handbook* shows, first, how policy evaluation has been institutionalized worldwide and within political systems. Then, four chapters cover the specific role of policy evaluation for national parliaments, public administrations, courts, and public deliberation during voting campaigns. The subsequent chapters move to the international level, examining the role and functioning of policy evaluation in the European Union and in international organizations as well as the way to build evaluation capacity within non-governmental organizations and developing countries. Finally, this part of the *Handbook* intensively discusses the professionalization of evaluation and the pros and cons of a certification system for evaluators.

Over several decades, the practice of evaluation has spread around the world. In Chapter 12, Steve Jacob observes very different trajectories of evaluation systems across countries and aims at understanding the main drivers and effects of this differentiated evolution. He synthesizes the results from various international research projects, comparing and even ranking the institutionalization of evaluation across countries. Jacob argues that the main elements contributing to the construction of a national evaluation system include contextual factors focused on performance and accountability (see Chapter 1 by Patton, Chapter 6 by Peters and Pierre, and Chapter 7 by Han); the motivation and interests of political actors, decision-makers and other policy entrepreneurs; and, finally, the ability to build effective evaluation capacity throughout the administrative system (see also Chapter 14 by Kuhlmann and Veit). A presentation of the effects of institutionalization on the utilization of evaluation and good governance concludes this chapter.

Previous studies have recurrently indicated that policy evaluations have become increasingly important for parliaments and elected representatives. Yet most of these studies show that parliamentarians only use evaluations in a limited capacity. Pirmin Bundi thus proposes a new classification of evaluation use in parliaments to better account for this specific institutional context. His classification proposes two different dimensions: utilization rationale and parliamentary power (legislation vs oversight). Using the examples of three quite different parliamentary systems – those of the United States, the United Kingdom and Switzerland – Chapter 13 demonstrates that parliaments tend to use evaluations for oversight purposes independent from the country context. Bundi's findings suggest that the resources of parliamentary services must be increased to strengthen the systematic use of evaluation findings by elected representatives.

Chapter 14, by Sabine Kuhlmann and Sylvia Veit, addresses the role of evaluation *of* and *in* public administration. Proceeding from a broad definition of evaluation, which includes the variants of external and internal (self-) evaluation as well as ex ante, ex post and ongoing evaluation (see also Chapter 4 by Flückiger and Popelier), the authors focus on two key analytical dimensions: the provider of the evaluation and the subject of the evaluation. On this basis, four major types of evaluation are distinguished: (1) external institutional evaluation; (2) internal institutional evaluation; (3) external evaluation of administrative action/results; and (4) internal evaluation of administrative action/results. Types 1 and 2 refer to evaluation *of* administrative structures and processes as the subject of administrative reform. Types 3 and 4, by contrast, represent different versions of evaluation *in* public administration, because the subject is administrative action and its outputs. The chapter highlights salient approaches and organizational settings of evaluation and provides insights into the institutionalization of an evaluation function in public administration. Furthermore, it explores concrete examples to illustrate the different types of evaluation of and in public administration. Finally, Kuhlmann and Veit draw lessons regarding strengths and potential but also remaining weaknesses and challenges of evaluation of and in public administration, particularly addressing the embeddedness of evaluation in political processes and the crucial issue of knowledge/evaluation utilization.

Evaluation research addresses state action directly, hence its suitability for challenging or justifying policies in place. How do courts benefit from it? This is the question Simon St-Georges and Christine Rothmayr Allison focus on in Chapter 15 (see also Chapter 4 by Flückiger and Popelier). They first propose a literature review and a framework for the use of evaluation in court. Evaluation results might point to unintended side-effects or highlight state failure and be used for antiregulatory reforms. To the contrary, they might also be suitable for defending policies by refuting assumed causality between state action and possible negative impact or for reinforcing and modifying policies. Applying key concepts on evaluation use, the authors then analyse a sample of court cases in various social and environmental fields in Europe, North America and worldwide to investigate whether judges relied on evaluation results and under what circumstances evaluations were successfully used to challenge or support policies. St-Georges and Rothmayr Allison conclude with the following observed trend in evidence-based judicial review: ‘courts are more likely to engage with the absence of evaluations at the time of enactment, or when reviewing the process of ex ante assessments, rather than debate the results of substantial ex post policy evaluations’. This finding has practical implications for upcoming research design on the use of evaluation by courts.

Evaluation findings are not only relevant to elected representatives, civil servants or justices, as discussed previously. They can also influence how ordinary citizens cast their vote if a popular vote is organized on a specific policy issue. Fritz Sager, Caroline Schlaufer and Iris Stucki address the question of the pertinence of evaluation findings for direct democracy. They refer to the broader literature on knowledge and democracy and employ data from a large Swiss research project to make four arguments. First, voters are interested in evaluation results. Second, evaluation use makes a difference in the quality of democratic discourse. Third, evidence alone does not conquer hearts; the use of evaluation results increases trustworthiness but not the emotional appeal of an argument. Fourth, despite its beneficial effects, voters lack accessible evaluation-based information in direct-democratic campaigns. Based on these innovative insights, the authors conclude by challenging pessimistic normative accounts on the relationship of (evaluation) knowledge and (direct) democracy.

In Chapter 17, Paul J. Stephenson and Jonas J. Schoenefeld discuss the role and functioning of policy evaluation in the executive and legislative venues of the European Union (EU). The EU's Better Regulation agenda has placed more importance on the role of ex post policy evaluation to close the policy cycle and to provide lessons for policy design and drafting of new legislation (see also Chapter 3 by Fontaine). Recent innovations include the 'evaluate first' principle, the European Commission's Regulatory Fitness and Performance programme (REFIT), fitness checks, and the Regulatory Scrutiny Board. The purpose and function of evaluation in the European Commission have evolved over 40 years across the Directorates-General. Today, the authors observe an increasingly harmonized approach to managing the evaluation cycle and processing findings from ex post evaluations. The European Court of Auditors is also active in auditing policy performance, examining the economy, efficiency and effectiveness of interventions. The European Parliament deliberates upon evaluation through its scrutiny and oversight work; in recent years it has increased its capacity to process and use evaluation, while also placing greater focus on assessing the performance of the EU laws and policies it co-legislates. The political, institutional and policy implications of ex post evaluation thus constitute a core area of interest for practitioners and researchers of the EU.

Is this encouraging trend of evaluation development in the EU also at work in international organizations (IOs)? Chapter 18 by Valentina Mele delivers instructional answers to this question. Owing to their nature as information brokers with a comparative perspective, IOs offer a natural venue to better understand which policy interventions work and how. In recent years, specific procedures and ad hoc units to advance the practice and culture of evaluation in IOs have rapidly grown despite the minimal scholarly attention they have received thus far (see also Chapter 12 by Jacob). Mele takes stock of existing literature and presents how policy evaluation is enacted by IOs, offering examples from current institutional developments. She positions IOs as autonomous policy actors that employ evaluation as a strategic governing tool. She reviews some of the main conceptions and definitions of policy evaluations put forward by IOs, identifying their common denominator and variations. Mele observes that IOs strive to shield the politically sensitive evaluation process from undue political influences by establishing autonomous units; however, such efforts are fraught with contradictions. They build policy evaluation capacity by decentralizing the evaluation practices in the countries, typically in the guise of a local partnership with national institutions and professionals; by ensuring direct support; and by establishing global evaluation networks that operate as epistemic communities in which experts frequently interact and develop joint problem definitions and solutions. IOs also set evaluation criteria, benchmark countries and other stakeholders against those criteria, and disseminate the results (see also Chapter 21 by Gauthier and Roy). Such purportedly neutral techniques are de facto charged with policy purposes and represent crucial instruments of governance.

The inherent strategic and political dimension of policy evaluation can also be highlighted when evaluation is in the hands of philanthropists and non-governmental organizations (NGOs). David J. Gilchrist and Ben Perks stress that philanthropy is an important source of NGO funding, whether as a top-up to publicly sourced funds or as a primary resource. While philanthropy might have previously provided funds for capital projects, it has increasingly sought to impact broader public policy development and evaluation. In a similar vein, NGOs also seek to impact policymaking in the interests of constituents. Chapter 19 thus examines the phenomenon of philanthropist and NGO human services policy evaluation in the context of the political pressures in the Australasian region. The authors focus notably on accountability mechanisms (see also Chapter 7 by Han) and analyse to whom and of what philanthropic foundations have to give an account. These issues are particularly relevant in the Australasian funding environment, which is marked by new public management allocation methods (see also Chapter 6 by Peters and Pierre). Gilchrist and Perks eventually provide prescriptions for effective and politically resilient policy evaluation as well as organizational learning processes.

Chapter 20 discusses the development of policy and programme evaluation by governments in low- and middle-income countries in Latin America, Africa and Asia. Ian Goldman, Thania de la Garza Navarrete, Asela Kalugampitiya, Alonso Miguel de Erice Dominguez, Edoé Djimitri Agbodjan, Takunda Chirau and Ayabulela Dlakavu provide a framework for understanding evaluation systems and describe what elements of evaluation systems can be found in the aforementioned three regions. Evaluation systems differ from country to country depending on the context, motivation, need and demand (see also Chapter 12 by Jacob). However, relatively few countries have well-developed evaluation systems, systematically evaluate their policies, and use their evidence to inform programme design, planning and budgets. The authors claim that a legal foundation for the evaluation systems would help make them more robust in the face of political and administrative changes. Finally, complementing the preceding chapter on philanthropic foundations and NGOs, Goldman and his colleagues address the role of donors in supporting evaluation in low- and middle-income countries. In particular, they explore recent research into the value and uses of evaluation and outline challenges faced when establishing an effective evaluation system, such as internal capacity-building instead of external evaluation by international experts and the departure from a hierarchical and punitive bureaucratic culture (see also Chapter 5 by Dunlop and Radaelli).

The last chapter in this part (Chapter 21) addresses ‘evaluation professionalization’ as a challenge that all national evaluation systems and supra- or international institutions face once they have achieved a certain level of maturity. Benoît Gauthier and Simon N. Roy define professionalization as the process of transformation of a trade from a loose artisanship to a codified and regulated occupation. Furthermore, the authors identify four successive phases of this process: unstructured occupation, voluntary and structured collective, formal professional recognition, and regulation and licensing. They clarify the dynamics at play in the mutation and weigh the advantages and disadvantages of professionalization. They also investigate the tools that support it at each phase of the process, including norms of practice, ethical guidelines, competency frameworks and recognition systems, as well as the conditions for burgeoning in the ecology of national evaluation systems (see also Chapter 12 by Jacob). These conditions include formal and informal demand for and existence of evaluation expertise, a training system, and attention of civil society to government performance. Lastly, Gauthier and Roy suggest that IOs (see also Chapter 18 by Mele) are one key actor supporting the professionalization of policy evaluators, as illustrated by the leadership of the United Nations Evaluation Group.

PART IV: EVALUATION AND BEHAVIOURAL PUBLIC POLICY

The fourth part of the *Handbook* aims to bring recent trends in behavioural public policy into the literature on policy evaluation. It scrutinizes whether evaluation is reinforced by the emergence of ‘nudge units’ across the world, which systematically use randomized controlled trials, and by field experiments to improve public service delivery and citizen co-production of public policies. It also addresses the transformation of policymaking and evaluation practice through the ongoing digitalization trend and the use of AI solutions and services. Finally, two chapters are dedicated to the impacts of street-level bureaucrats' behaviour on policy outputs as well as the intriguing 'non-take-up' phenomenon (i.e., eligible citizens not asking for social benefits they are entitled to receive) that is frequently observed in the evaluation of welfare policies.

Chapter 22 by Peter John reviews the historical development of randomized controlled trials (RCTs) and behavioural change policies (behavioural insights [BIs]) as the 'twins of modern policy evaluation'. John notes how trials moved from complex evaluations, usually of social policies, to more rapid and generic testing, aided by the growing popularity of BIs, using examples from the United Kingdom and the United States, Europe, as well as Australia. Concurrent with this growing popularity of robust evaluation, challenges to scientific hegemony have emerged through movements that question the authority of facts and give great value to intuition and popular feelings about policy, leading to what is called the 'post-truth world' (see also Chapter 1 by Patton). Also, BIs may be seen as essential liberal paternalists in that they are used to decide policies for people and then seek to manipulate them, using RCTs to achieve a predetermined outcome decided by scientific experts. This appears to be just the kind of policy that populist leaders would seek to resist, but evidence suggests continuing popularity of such nudge policies (Sunstein, 2020). One answer to this paradox is that nudges and the use of RCTs are probably less top-down than they commonly appear and imply policymaking informed by trial and error. Many nudges seek to encourage (slow) reflection by citizens rather than rely on (fast) automatic decision-making processes. The 'nudge plus' programme that is suggested by John aims to develop these reflective devices and thus legitimate both BIs and RCTs in a post-truth world.

As a matter of fact, field experiments in which some organizations or citizens are randomly assigned to one policy and others are not, are increasingly used as a methodological tool to evaluate policy effects. In Chapter 23, Simon Calmar Andersen reviews state-of-the-art knowledge about how and when to use field experiments. First, the method of field experiments is briefly presented to demonstrate why they are often seen as the gold standard in estimating the 'causal effects' of a public policy. Second, Andersen introduces key concepts in a discussion of how the knowledge gained from field experiments regarding specific policies can be generalized to similar policies in other contexts? Finally, the chapter considers prospects and challenges. Several ethical questions are raised regarding the active manipulation by evaluators of policies affecting people or organizations: when is this manipulation unethical, and how does it differ from other methods of empirical research on public policy? Examples from existing field experiments (e.g., education policy in Denmark) are used to illustrate these dilemmas.

Tereza Cahlikova and Omar Ballester note that public policy evaluation has traditionally been conducted in a context constrained by legal, institutional and political red tape. However, the arrival of digitalization and applications of AI add a new layer to public policy evaluation processes. What does this disruptive technological innovation change for the evaluation process and for the evaluators? Assessing the quality of technologies, tools and processes in use requires that evaluators revise their skillset and methods. Cahlikova and Ballester aim to illustrate the challenges that digitalization entails for policy evaluation. First, they examine how the evaluation of digitalization-related public policies differs from evaluation of their 'analogue' counterparts using empirical findings from a Swiss local government project that introduced digital education in primary and secondary schools. Second, they discuss how the responsible implementation of (automated) decision-making assisted by AI should be conducted and how AI impacts practical evaluation (see also Chapter 4 by Flückiger and Popelier). Some governments are aware of this ardent challenge and have started imposing an 'algorithmic impact assessment' (e.g., the Directive on Automated Decision-Making issued in 2019 by the Government of Canada). Chapter 24 thus contributes to the burgeoning discussion of the prospects of policy evaluation amidst the proliferation of digitalization and AI in the public sector. It raises important questions about 'autonomous learning' as a new type of learning influencing and resulting from policy evaluation (see also Chapter 5 by Dunlop and Radaelli).

In measuring and valuing the effects of policies on the ground, evaluations should, almost by definition, consider the crucial role of street-level bureaucrats (SLBs). SLBs represent the front lines of government policy: they interact directly and recurrently with the policy target groups and use their discretionary power to implement policy instruments (Lipsky, 1980). Understanding SLBs' behaviour is thus key to explaining the delivered policy outputs vs implementation gaps (see also Chapter 26 by Daigneault) and the subsequent policy outcomes. More specifically, analyses of SLBs in policy evaluations shed light on what happens at the level of individual SLBs and how politics continue at the front line of implementation. Considering the remarkable attention placed on behaviour in policy and public administration studies, Eva Thomann and Eva Lieberherr analyse how the behaviour of SLBs influences the implementation of policies and, thereby, contributes decisively to policy effects. Based on a literature review, they present descriptive concepts including policy alienation and enforcement styles; explanatory factors such as heuristics and cognitive biases of SLBs; and accountability relations between SLBs and other policy actors (see also Chapter 6 by Peters and Pierre, and Chapter 7 by Han). Implications of SLBs' behaviour for the evaluation of public policies are discussed. In these ways, Chapter 25 provides insight for policy evaluators by introducing the key concepts to consider when assessing the role of SLBs in making policies work in practice.

The situation of non-take-up (NTU), which occurs when individuals do not receive the public services or social benefits to which they are formally entitled, is one output of the interaction between SLBs and policy target groups. Non-take-up is indeed a serious policy problem that has deleterious consequences for individual citizens, marginalized social groups, as well as policy effectiveness (van Oorschot, 1995; see also Chapter 10 by Pires and Lotta). Moreover, NTU poses significant challenges when designing, implementing and evaluating public policy. Pierre-Marc Daigneault thus draws on the insights from the literature on public administration, economics, sociology, social work and policy evaluation to investigate these policymaking challenges. Looking at NTU through a public policy lens, the author explains how to measure empirically and evaluate normatively programme (non-)take-up. A three-stage framework (i.e., threshold, trade-off and application stages) and the concept of administrative burden structure the review of NTU determinants. Daigneault concludes by examining prospects pertaining to the evaluation of programme (non-)take-up. He suggests that policy evaluation should contribute to the fight against NTU, since it has detrimental effects on citizens, policymaking and democracy.

OUTLOOK

The aim of this *Handbook* is to put public policy evaluation into its policymaking and democratic context. The various contributions point to several challenges policy evaluators face in their daily practice. Indeed, evaluators are always embedded in a singular political context, and one evaluation study cannot simultaneously serve all potential purposes that could ideally be expected from a policy evaluation process. These include bureaucratic accountability about policy implementation; instrumental learning from innovative policy experiments; conceptual enlightenment about the causal mechanisms underlying a policy intervention; empowerment and social emancipation of marginalized groups; and provision of arguments to fuel the public debate, among others (Chelimsky, 2006).

The authors of the 26 *Handbook* chapters address these topics from different but complementary standpoints. We hope that the readers will enjoy navigating across different epistemological, theoretical, methodological, but also normative perspectives. In his inspiring introduction to a recent volume (also published by Edward Elgar Publishing) about the future of evaluation research, Peter Dahler-Larsen (2021, p. 9) stresses that '[a]ll forms of evaluation depend on some starting point; some distinction between what is questionable and what is less questionable, what is contestable and what is not. In democracies, of course, everything can be debated. But if evaluation is a debate and only a debate, it is not an evaluation'.

The present *Handbook* does not privilege one standpoint over another. This would not make sense, since each evaluation process is contingent. On the contrary, this *Handbook* is conceived as a thesaurus to consult for further thinking and critically reflecting about one's own evaluation practice and its impacts on policymaking processes and democracy. Most chapters are illustrated through concrete evaluation examples from countries across the globe, at different levels of power (i.e., subnational, national, international and supranational institutions), and encompassing a large diversity of policy domains. Of course, this broad scope does not do full justice to all practical activities and academic debates within the evaluation field. However, it should help those readers considering how to best engage in policy evaluations with the noble objective of improving policymaking in democracies.

REFERENCES

Barrados, M., & Lonsdale, J. (Eds.). (2021). *Crossover of audit and evaluation practices: Challenges and opportunities*. Routledge.

Bovens, M. (2007). Analysing and assessing accountability: A conceptual framework. *European Law Journal*, *13*, 447–468.

Bundi, P., & Trein, P. (2022). Evaluation use and learning in public policy. *Policy Sciences*, *55*, 283–309.

Campbell, D. T. (1991). Methods for the experimenting society. *Evaluation Practice*, *12*(3), 223–260. (Original Work published 1971).

Chelimsky, E. (2006). The purposes of evaluation in a democratic society. In L. Shaw, J. C. Greene & M. M. Mark (Eds.), *The SAGE handbook of evaluation* (pp. 35–55). SAGE.

Christie, C. A. (2007). Reported influence of evaluation data on decision makers' actions: An empirical examination. *American Journal of Evaluation*, *28*(1), 8–25.

Dahler-Larsen, P. (2021). Introduction to *A Research Agenda for Evaluation: Inspirational Themes*. In P. Dahler-Larsen (Ed.), *A research agenda for evaluation* (pp 1–14). Edward Elgar Publishing.

Eliadis, P., Furubo, J.-E., & Jacob, S. (Eds.). (2011). *Evaluation: Seeking truth or power?* Transaction Publishers.

Fischer, F. (2005). *Evaluating public policy*. Wadsworth (Original work published 1995).

Fischer, F., & Forester, J. (1993). *The argumentative turn in policy analysis and planning*. Duke University Press.

Guba, E. G., & Lincoln, Y. S. (1989). *Fourth generation evaluation*. SAGE.

Hill, M., & Varone, F. (2021). *The public policy process*. Routledge.

House, E. R. (1972). The conscience of educational evaluation. *Teachers College Record*, *73*(3), 405–414.

Howlett, M., Ramesh, M., & Perl, A. (2009). *Studying public policy: Policy cycles and policy subsystems*. Oxford University Press.

Lasswell, H. D. (1936). *Politics: Who gets what, when and how*. Meridian Books.

Lipsky, M. (1980). *Street-level bureaucracy: Dilemmas of the individual in public services*. Russell Sage Foundation.

Organisation for Economic Co-operation and Development (OECD) DAC Network on Development Evaluation (n.d.). *Evaluation of development programmes*. https://www.oecd.org/development/evaluation/.

Pawson, R., & Tilley, N. (1997). *Realistic evaluation*. SAGE.

Power, M. (1997). *The audit society*. Oxford University Press.

Scriven, M. (1991). *Evaluation thesaurus*. SAGE.

Sunstein, C. R. (2020). *Behavioral science and public policy*. Cambridge University Press.

van Oorschot, W. (1995). *Realizing rights*. Avebury.

Weiss, C. H. (1972). Evaluating educational and social action programs: A ‘treeful of owls’. In C. H. Weiss (Ed.), *Evaluating action programs* (pp. 3–27). Allyn & Bacon.

Weiss, C. H. (1975). Evaluation research in the political context. In E. L. Struening & M. Guttentag (Eds.), *Handbook of evaluation research* (Vol. 1, pp. 13–26). SAGE.

Wholey, J. S., Scanlon, J. W., Duffy, H. G., Fukumotu, J. S., & Vogt, L. M. (1970). *Federal evaluation policy: Analyzing the effects of public programs*. Urban Institute.