Policy evaluation styles

De Francesco, F.; Pattyn V.; Howlett, M.; Tosun J.

Policy evaluation styles

Fabrizio De Francesco and Valérie Pattyn

The development of evaluation as a profession is marked by different diffusion waves and assumptions of what good governance and associated evaluation evidence entail. With every new wave, an increasing number of countries adopted common evaluation practices. The question is whether this also implies commonalities in the way that evaluations are practised nowadays, or whether differences rather prevail.

In this contribution, we analyse the volume and type of sediments which different waves have deposited. We focus on three evaluation practices that represent the major waves and that are exemplary for regulatory governance in Europe: administrative burden measurement or standard cost model; regulatory impact assessment; and randomized controlled trials and nudges.

We contribute to both the policy evaluation literature and the literature on policy styles by proposing a typology that links waves of evaluation diffusion with prescriptive evaluation theories. While we do not find evidence for a homogenisation of evaluation styles, we highlight that the practices based on economic methods for consolidating evidence have a substantial impact on public administration culture.

Keywords: Administrative burdens; behavioral insights; European Union; regulatory governance; policy appraisal, policy evaluation

1 Introduction

Since the 1960s, with the emergence of the evaluation field, policy evaluation has been in constant flux. New evaluation approaches were introduced, others were heavily criticised and disappeared or regained momentum in revised ways (Furubo and Sandahl, 2002; Vedung, 2010). Full of many techniques for assessing policy effectiveness and efficiency, the current toolbox is the result of sedimentation from different waves of evaluation diffusion (Furubo and Sandahl, 2002; Stame, 2003; Vedung, 2010). Each wave came with specific assumptions of what good governance and associated evaluation evidence entail. Since every new wave also contributed to the dissemination of evaluation practices across European Union (EU) countries (Jacob et al., 2015; Pattyn et al., 2018; Stockmann and Meyer, 2020), the aim of this contribution is to observe the extent of commonalities in the way that evaluation is practiced nowadays. On the one hand, one could argue that with each wave novel but largely diffused methods are laid over existing and country-specific evaluation approaches, gradually resulting in an 'international evaluation culture' (Barbier and Hawkins, 2012: 12). Countries can be assumed to develop isomorphic behavior in this regard (Dimaggio and Powell, 1983). This converging trend would be in line with the efforts of international organisations (such as the EU, the World Bank, and the OECD) to promote common evaluation standards and techniques (Furubo and Sandahl, 2002; Stame, 2003). On the other hand, despite this extensive process of homogenization, evaluation practices may still differ across countries. Whether and how certain evaluation methods are adopted is to large extent mediated by a given country's administrative and legal tradition (Peters, 1997; 2008).

Evaluation is a political activity by nature, and its contours happen in the context of polities and political communities (Weiss 1993; see also Barbier and Hawkins, 2012). Therefore, an evaluation style, defined here as the evaluation practice consolidated in a particular country's administrative tradition, influences the policy formulation, and the resulting outputs. However, evaluation styles may be altered by the consolidation of methods and practices brought by successive waves. The speed of convergence towards a common international evaluation culture depends on whether the sediments from different diffusion waves are large or rather small.

Testing convergence in evaluation styles constitutes a challenging empirical puzzle. And we are not the first in addressing it (see e.g. Barbier and Hawkins, 2012 for a country by country volume on 'evaluation cultures'). However, our approach is unique. By analyzing the extent of institutional change brought by three regulatory appraisal systems, we differentiate how successful each wave is nurturing a common evaluation culture. Specifically, we look at countries' experience in (1)

administrative burden measurement or Standard Cost Model (SCM), as the main sediment of the new public management and neo-liberal wave, (2) regulatory impact assessment (RIA), that has been brought in many countries by both the neo-liberal wave and the evidence-based policy wave (Radaelli 2010), and (3) randomized controlled trials (RCTs) to test "nudges" and behavioral insights, that are predominantly associated with the most recent evidence-based policy-making wave. The evaluation methods selected share the purpose of enhancing regulatory quality and therefore allows for a meaningful comparison. Empirically, although we acknowledge that evaluation practices spread also across policy fields (Barbier and Hawkins; 2012; Pattyn et al., 2018; Stockmann and Meyer, 2010), our level of analysis is at the country level. And we limit our observations to EU countries, allowing us to magnify the convergence towards a common evaluation styles.

In the next section, we concisely review different possible conceptions of evaluation style and link different waves of evaluation diffusion with prescriptive evaluation theories. This conceptual map sets the background for the actual empirical analysis in which we approach evaluation styles from a triple lense. We first identify how countries define the different regulatory governance practices. Second, we describe to what extent SCM, RIA and nudge units have spread, shedding light on their diffusion drivers and barriers. Finally, we look at the actual implementation and institutionalisation of the practices. In the conclusion, we discuss whether our qualitative findings confirm the emergence of a common European evaluation culture.

2 Evaluation styles? A conceptual exploration

A plethora of conceptualizations of policy evaluation circulate in the literature (King, 2003), which is partially due to the strong practitioner-oriented nature of the field, with different evaluators often having different understandings of what an evaluation is. Inspired by Vedung (2010), we opt for a deliberately parsimonious and comprehensive definition of evaluation: a 'careful assessment of public sector interventions, their organization, content, implementation, outputs, or outcomes, which is intended to play a role in future decision situations'.

Several scholars attempted to develop taxonomies of evaluation approaches (see e.g. Shadish et al., 1991; Widmer and Rocchi, 2012). The most seminal one is probably Alkin and Christie's (2012) evaluation tree. They grouped all main evaluation prescriptive theories in three branches depending on whether they put main emphasis on the methods for conducting evaluations, the valuing aspect of evaluations, or the use of the evaluations. We maintain that each evaluation branch corresponds to a particular evaluation style.

From a more historical perspective, evaluation styles can be approached along the lines of five different generations. Table 15.1 provides an overview of the distinguishing characteristics of each generation. The last three generations still reflect the main approaches around which the present day evaluation field revolves. Typically associated with scholars as Scriven (1991), and Rossi and Freeman (1985), the third generation is still considered the mainstream evaluation approach. The evaluator takes up the role of 'judge', which differs from the mere technical and descriptive roles of the previous two generations (Guba and Lincoln, 1990, p.30). By relying on effectiveness and efficiency criteria, the evaluator is supposed to make explicit judgements on the merit or worth of the evaluand. The 'fourth generation', introduced under this very label by Guba and Lincoln (1990), represents a responsive social constructivist approach to evaluation. This evaluation generation is an antonym of 'preordinate' evaluation where the parameters and boundaries of the evaluation have been defined prior to the start of an evaluation. The fourth generation is characterized by an "an interactive, negotiated process that involve all stakeholders" (Guba and Lincoln 1990, 38-39).

Scholars recently identified the emergence of a fifth generation (Brousselle and Buregeya, 2018), coined as the 'explanation' generation. This generation emphasizes the importance of theory-based approaches to evaluation such as realist evaluation (Pawson and Tilley, 1997), contribution analysis (Mayne, 2012), and logic analysis (Rey et al., 2012). These methods highlight the explanatory power of contextual characteristics, implementation processes, and causal pathways to show how an intervention's activities and outputs led to outcomes.

Generation	1st: Measurement	2nd: Description	3rd: Judgement	4th: Pluralism	5th: Explanation
Role of evaluator	To provide technical expert	To describe	To judge	To negotiate	To explain
Key characteristics	Test to measure effectiveness	Objectives; Formative evaluation	Standards, Objective evaluator	Primarily qualitative methods; No causally	Theory-based approaches

				inferential statistics	

Table 15.1: The five generations of evaluation (Authors, elaborating on Lai, 1991, p.4)

Instead of generations, others used the diffusion waves analogy to describe different fads and fashions in evaluation practicing over time. Stame (2003) and Furubo and Sandahl (2002) distinguish three major waves of evaluation diffusion, each relying on different catalyzing factors. The first wave mainly originated in the US with the Great Society (1960s-1970s), but also swept over some European countries as Sweden, the Netherlands or the United Kingdom. Evaluation requirements were structurally incorporated in the introduction of new social programs and political tools, and reflected a political culture favoring an engineering or rationalist approach to policy making. Those pioneering countries in policy evaluation all invested tremendously in the development of strong national statistical institutions (Furubo and Sandahl, 2002). The second wave of diffusion (1980s-1990s) is marked by the launching of the New Public Management (NPM) doctrine, first in Australia and New Zealand, but also in some European countries. The creation of independent agencies and quangos had a major impact on the rowing and steering capabilities of ministers. Measures of efficiency and effectiveness became a core prerequisite for the well-functioning of NPM modeled administrations. Since the 2000s, the third wave is associated with the influence of international organisations such as the European Union and the World Bank on evaluation practice. The structural funds program, for instance, has played a major role in the spreading of evaluation practice across European countries (Stame, 2003; Pattyn et al., 2018).

Also Vedung (2010) identified different evaluation waves. However, he is more explicit about the type of approaches and methods they are commonly associated with. As such, he marked the 1960s as the 'scientific wave', with evaluations to be conducted by academics, preferably via RCTs, to identify the most effective means to an end. This wave eroded in the 1970s, with the emergence of more democratic, reponsive and participatory-based types of evaluation, coined under the label of the 'dialogical wave'. This wave brought about the importance of users and practitioners' perspectives on public service and policy quality. The 1980s saw the dominance of NPM managerial thinking. This 'neo-liberal wave' puts an emphasis on efficiency-oriented (such as cost-benefit analysis; cost-effectiveness analysis) and client-oriented approaches to evaluation. More recently, the 'scientific wave' has returned, though now framed under the umbrella of 'evidence-based policy making' (EBPM) and 'what works'. Method-wise, this wave involves a renaissance

for scientific experimentation and RCTs, often associated with behavioural insights and nudging. Meanwhile, consensus is growing, at least in evaluation discourse, that no single evaluation approach has the monopoly on good evidence, and that different approaches to causality are fit for different policy settings (Pattyn, 2019).

Having sketched the different generations and diffusion waves, it is useful to link them to the particular branches of Alkin and Christie's theory tree. As such, in Table 15.2 we combine the strengths of both a historical as well as a more theoretical outlook on evaluation practices. With this chapter focusing on evaluation practices of regulatory governance, we do not put too much emphasis on the dialogical wave that preceded the emergence of the regulatory state in Europe and elsewhere but the United States (Majone 1994). Further, the NPM and the EBPM waves epitomise the evolution of evaluation practice within the European Commission. In Table 15.2, the column 'methods' describes several dimensions and features of the method branch of the evaluation tree. This branch is linked with the scientific mode of evaluation. This mode is pursued through scientific methods as RCTs and quantitative analysis for establishing cause-effect relationship. Accordingly, it is predominantly associated with the first (measurement) and the more recent (explanation) generation. The 'valuing' evaluation branch is mainly associated with the NPM wave, given its emphasis on value for money and customer satisfaction. Through the development of professional evaluation standards typical of the third generation, evaluators, consultants, and auditors are given a central role in this regard.

Features	Evaluation branches	
	Methods	Valuing
Social context and wave of diffusion	Scientific & EBPM waves	NPM & neoliberal wave
Scope	Best scientific evidence; focus on understanding cause- effect of interventions	Making judgment; focus on accountability
Main actors	Scientists	Evaluators, Consultants, and Auditors
Practice	Experiments and quasi-experiments; systematic literature review and meta-analysis	Indicators, performance measures, ranking and benchmarking; Comparison and assessment of options

Research method	Quantitative models for cause-effect relationships	Indicators and quantitative measures for ranking and assessing performances and policy options
EU countries' pioneering experience	'What Works' centres and Nudge unit in the UK	Standard Cost Models in the Netherlands

Table 15.2: Features of regulatory evaluation branches in Europe

As mentioned in the introduction, the waves gradually implied a worldwide dissemination of evaluation practice, and triggered governments to further institutionalize evaluation in the policy process. Generally speaking, most EU countries have clearly matured in institutionalizing evaluation practice. Nonetheless, as systematic country comparions revealed (Jacob, Speer and Furubo 2015; Stockmann and Meyer, 2020), there remain substantial differences in the extent and modalities of institutionalization in the evaluation conducted both in the executive and legislative branches of government. Based on their cross-country comparions, Jacob et al. (2015) conclude that there are different paths towards evaluation culture. This led them to assume that '[n]ational policy styles can shape patterns of policymaking in systems of public administration, and it can be assumed that some of these national characteristics have an impact on evaluation regardless of the particularities of different policy fields and organizations' (Jacob, Speer, and Furubo 2015: 7). Intimate relationships exist in the way how policy evaluation is thought of, and organized within the institutional setting of a given country (Barbier and Hawkins, 2012: 6). It is not clear, though, to what extent this also implies that there are clear differences in using specific evaluation instruments brought by different evaluation waves. This is exactly the empirical question that we will discuss in the remainder of the chapter.

3. Convergence and divergence in regulatory evaluation practices

3.1. Administrative burden measurement and Standard Cost Model

Since the introduction of the NPM doctrine, central emphasis has been put on the reduction of regulatory administrative burdens or red tapes. 'Administrative burdens refer to regulatory costs in the form of asking for permits, filing out forms, and reporting and notification requirements for the government' (OECD 2006, 9). A broader definition is provided by Burden et al. (2012, p. 741) who conceives administrative burden as 'an individual's experience of policy implementation

as onerous'. Sustained when firms and citizens interact and exchange information with the public administration, administrative burdens can entail learning (the search processes to collect information), psychological (associated with individuals' emotional state for dealing wth administrative processes), and compliance costs (the costs of completing forms and providing documentation of status) (Moynihan, Herd and Harvey, 2015). It is clear that the OECD's definition and countries' effort to reduce administrative burden is limited to the narrow element of the latter category of compliance costs, as such overlooking the other (and possibly more relevant) elements of administrative burdens.

The most popular evaluation tool for reducing administrative cost, the Standard Cost Model (SCM) has particularly been engrafted on this conception. SCM is 'a policy instrument for measuring the compliance costs of legal information obligations from businesses, institutions and civilians to government and governmental institutions' (Nijsen 2013). As a cost minimization technique, it relies on a relatively simple measurement formula: The cost of each information obligation incurred by a 'typical firm' is multipled by the number of firms and the times the information is required (Torriti, 2007) . The cost estimation may rely on surveys of a representative sample of firms. Although the SCM has been used for ex-post evaluation of business regulatory environment, the tool is also exploited for assessing the likely impact of administrative burdens. Indeed, granted that regulatory benefits remain unquestioned by the SCM, the quantification of administrative costs allows to set reduction cost targets (Coletti and Radaelli 2013) and score regulatory options (Nijsen 2013).

Almost the totality of the EU member states currently applies SCM and related techniques (Wegrich 2009; Heidbreder, Wegrich, and Fazekas 2010). With the potential to enhance business competitiveness and the efficiency of a country's regulatory and administrative environment (Coletti and Radaelli 2013), the politically neutral cost-effectiveness (to attain the desired goals and social benefits) methodology has been the main driver of diffusion (Nijsen 2013). This rapid diffusion has been facilitated by the simplicity of the underpinning evaluation methodology that lends itself to relatively easy implementation (Nijsen 2013, p. 234). In other words, '[i]t appears as if the glow of pragmatism rests on SCM and it is this, which made it so popular particularly among government officials' (Weigel 2008, p. 26). With its narrow concept and simplicity, it has been a methodology that could be universally adjusted to the rulemaking and evaluation practice of each country.

In Europe, the adoption of SCM has been advocated by the International SCM Network (Nijsen 2013), a network of high-level civil servant led by the Dutch government that has now ceased its activities. By agreeing on a common manual for measuring administrative burdens, the International SCM Network was also able to set the agenda in several international organisations. In mid-2000s, both the World Bank and the OECD initiated an international reviews of the SCM implementation (OECD 2006; World Bank 2007). While the reviews exhibited some skepticism of the evaluation methodology, it can also be read as a legitimisation of the method by international organisations. In fact, the reduction of administrative burdens was mainly promoted as an initial step towards more ambitious reforms of regulatory governance and as a feasible solution for countries with limited administrative capacity (Coletti and Radaelli 2013; Nijsen 2013). Also the European Commission played a major role in the diffusion through the lauch of the 2007 action programme to reduce administrative burdens by 25% by 2012. Whereas before 2007, only 10 EU member states adopted the SCM, at the end of 2009, Slovakia was the only country that did not adopt any administrative burdens reduction strategy (Heidbreder, Wegrich, and Fazekas 2010).

Its straightforward applicability and the promotion by international organisations and pioneering countries as the Netherlands contributed to a relatively uniform administrative burdens' evaluation style in Europe (De Francesco 2011). This is not to say that there is no cross-country variation in other dimensions of institutionalisation, such as the establishment of an oversight body that ensures the accomplishment of the reduction target, the baseline measurement for the stock of regulation, and the establishment of an administrative burdens reduction target. However, this variation is largely associated with the time necessary for institutionalising such regulatory governance instrument (Heidbreder, Wegrich, and Fazekas 2010). It is not surprising that pioneers and early adopters have implemented the SCM extensively vis-à-vis later adopters. Therefore, we have no indication that the implementation of SCM is strongly associated with broader institutional conditions and countries' particular policy style.

Although SCM is still widely used today, which can be seen as a clear sediment of NPM, the methodology has been increasingly contested by scholars, mainly for its limited economic rationality (Torriti 2007). Criticism especially increased, since the EBPM wave reached Europe in the 2000s, and with the 2008-9 global financial crisis. The constellation of the financial crisis and other wicked problems such as climate and technological change forced European governments and the EU institutions to go beyond internal management problems, and to consider 'what works' from a broader cause-effect point of view. Thus the international and national political agenda supporting the SCM and the administrative burdens has consistently faded away. Indeed, since

2005 the World Bank has been publishing global rankings of administrative burdens, undermining the international promotion of the SCM and limiting the scholarly interest to the application of the administrative-burden effectiveness methodology to specific sectors such health and taxation (Nielsen et al, 2017).

3.2. Regulatory Impact Assessment

Regulatory impact assessment (RIA) revolves around a umbrella concept that is larger than the SCM. RIA encompasses 'a range of methods aimed at systematically assessing the negative and positive impacts of proposed and existing regulation' (OECD 1997). Although usually associated with cost and benefit analysis, Pareto efficiency and comprehensive rationality, RIA is frequently used to assess the impact of new regulations on business and social welfare, administrative and paperwork burdens, regulatory burdens on small businesses, and the consequences for international trade and employment. Its particular methodology and design vary depending on the policy objectives, and a given country's administrative process (De Francesco 2013). The following two elements are nonetheless common to any RIA system: i) an explanation of the specific need for regulation; ii) a systematic and consistent economic appraisal of foreseeable impacts arising from that regulation.

Since the end of the 1990s, RIA has been actively promoted by international organisations, initially as part of an NPM neo-liberal wave, but later also as an essential supporting tool for EBPM. The OECD has been successful in creating a community of reformers engaged in disparate policy reform agendas, such as deregulation, business competitiveness, and regulatory quality (Radaelli 2005). At present, it can be said that RIA has some kind of 'inherent symbolism' (Mossberger 2000). Not only is it a referential symbol, a way to name and make sense of the required regulatory reform, but it is also a condensation symbol which exemplifies governments' association of regulation with poor economic and social performance (De Francesco 2013).

Although the RIA label has been attached to a variety of functional purposes, which has led to a loosely bundled, even ambiguous concept, its symbolism has allowed the OECD to provide its member states with inferential shortcuts about the model they can emulate (De Francesco 2013). By reframing RIA as a tool to enhance the empirical evidence of decision-making, the OECD triggered governments in taking the decision to adopt RIA, even despite a lack of evidence of the impact of RIA on citizens' welfare (Cowen 2005). This led to a diffusion wave around 1995, when the OECD launched regulatory reform recommendations based on RIA. Another diffusion wave

was pushed in motion with the European Commission's adoption of RIA in the beginning of 2000s.

The diffusion of RIA is almost complete among EU and OECD countries: A 2017 OECD survey attests that that the almost totality of 44 RIA systems assessed (including the European Union,several EU member states, and non-OECD members) has adopted RIA (OECD, 2018). Only Latvia responded of not having written RIA guidelines. Furthermore, 35 systems have a body responsible for reviewing the quality of RIAs, the so-called regulatory oversight body (OECD, 2018).

Qualitative analyses corroborated the important role of the OECD in promoting RIA, regulatory oversight bodies, as well as evaluative practices of CBA in environmental policy. The OECD regulatory review was the main driver of RIA's adoption in Austria, France, Germany, and Italy. RIA as a tool for the economic competitiveness agenda also constituted the main rationale for its adoption in Greece, Portugal and in several Central and Eastern European countries (De Francesco 2016). Quantitative studies revealed the patterns of interaction among governments, that were facilitated by the OECD, and that provided them with a simple cognitive map for taking the decision to adopt IA (De Francesco 2012).

When it comes to the actual implementation and institutionalisation of RIA, similar patterns have been found than for SCM implementation: scholars revealed a negative correlation between years of adoption and the extent of implementation (De Francesco 2013; Radaelli and Meuwese 2009). Facilitated by the ideational role of the OECD, trends of convergence are mainly associated with the legal design and the evaluative practices of regulatory analysis, but not with regulatory oversight and regulatory quality indicators. Also legal origin has a role in explaining the variation in implementation. English legal origin countries have the highest implementation scores, followed by French and German legal origin countries. Scandinavian legal origin countries instead even lag behind the post-socialist countries (De Francesco 2013). Importantly, and different from the SCM, the sophistication of the RIA tool involves the risk of mere symbolic adoption. RIA practice has resulted in sometimes severe criticism of not conforming to evaluation quality standards. For example, it has been criticized that RIAs are usually carried out by the same public officials who are responsible for drafting legislation, and who are foremost interested in producing a legally sound regulation which fits into the existing legal fabric. It has been also said that the persons working on RIAs often lack the competencies to perform an actual impact analysis (Stockmann and Meyer, 2020).

Compared to SCM, the RIA evaluative models and practices were not transferred through government-to-government communication among countries. Whereas the SCM was a more easily transferable evaluative model, governments that were engaged with the institutionalization of RIA relied on their own direct experience, and which was shaped by the administrative context. Accordingly, we can conclude that a country's RIA style is characterized by the extent and quality of its institutionalisation, which is in turn influenced by the maturity of its regulatory state. For its overarching goal of improving regulatory governance, RIA is an instrument that has spread in punctuated diffusion waves. Differently from SCM, however, and despite the criticism, RIA still has a central place on political and scientific agendas of many countries.

3.3. RCTs and Nudges

After having lost some fad with the emergence of the dialogical and participatory wave of evaluation, RCTs has reconquered a firm place in the evaluation toolbox with the EBPM movement. At present, RCTs are by many still conceived to be the golden standard in evaluation. Resonating with a rational and positivist take on policy interventions, RCTs and experiments at large have the ambition of discovering what works best to achieve particular societal outcomes (Pattyn, 2019). The renaissance of RCTs is clearly associated with the resurgence of nudging or behavioral insights practices for policy making, and behavioral public administration in general. Broadly speaking, behavioral insights are defined as an 'inductive approach to policy making that combines insights from psychology, cognitive science, and social science with empirically-tested results to discover how humans actually make choices' (OECD, 2017). Experimentation and piloting provide a cost-effective way to test different policy scenarios on a small scale. As nudges significantly alter the behaviour of firms and citizens without affecting the freedom of choice (Thaler and Sunstein 2009), the literature on public policy and administration has started exploring the insight of behavioral science in order to generate theoretical predictions at the level of individual attitudes and behaviour that are often implicit and seldom empirically tested (James, Jilke and Van Ryzin 2017; Grimmelikhuijsen et al 2017). Still a prevailing issue is how to achieve not only policy efficacy and efficiency but also legitimacy of nudging. In this context, a conceptual framework has been put forward for assessing the extent of effectiveness, efficiency, as well as stakeholder support (Tummers 2019).

While not entirely novel, behavioural insights especially gained popularity and momentum as an evaluation practice in mid-2000 on both sides of the Atlantic pond. Both the Obama administration in the US and the Cameron coalition government in the UK were fascinated by the

possibility to have a 'soft touch' to regulation in order to solve market failures and reduce negative externalities. Thaler and Sunstein's (2009) global best-seller book played a major role in this respect. By dint of theoretical explanations and practical examples, the authors were able to transmit their insights and shook the economic theoretical foundations based on fully economic rationality of market operators and individuals. Nowadays, behavioral insights are no longer a mere fad, but are increasingly entrenched in policy making of many countries (OECD, 2017). The application of behavioral science to public policy is common across the globe. In 2018, there were at least 202 public entities relying on nudges (Afif et al. 2018), a number which probably increased in the meantime. In Europe, in a time span of eight years ten EU countries established a nudge unit: Austria, Denmark, Finland, France, Germany, Greece, Ireland, the Netherlands, Sweden, and –the pioneer- United Kingdom (John, 2019).

A quantitative analysis has shown that nudge units has been an Anglo-American phenonemon and tends to be associated with right-wing governments (John 2019), as behavioral insights are often seen as manipulative techniques (Tummers 2019) and an expression of neo-liberalism justifying state retreatment and private solutions to social problems (John 2018). Therefore, tense debate exists about the nature of nudges, with critics claiming that behavioral sciences give primacy to individual-level analysis, and do not sufficiently take the wider social and political relationships into account. Irrespective this discussion, cross-country research by Sunstein, Reisch and Kaiser (2019) about European citizens' attitudes on nudges revealed relatively wide approval. Citizen evaluations turned out to be relatively similar in most countries, except for Denmark and Hungary. General distrust or fear of government seems to matter in this regard.

International organisations are again a key promoter of nudges across the Europe, but not (yet) to the extent of RIA and SCM. The OECD promotes nudges through networking activities and a recently launched toolkit for guiding practitioners and policy makers in developing behaviorally informed interventions. The EU Joint Research Centre organizes similar networking initiatives, and monitors the application of behavioural insights in European countries. However, the main driver of nudging diffusion might have been the UK Behavioural Insights Team (BIT). As John (2018, p. 81) puts it, 'BIT has played a role in the international diffusion of behavioural insights as used by government and agencies across the world, through its extensive network of international contacts and where other governments have seen BIT as a model to emulate'. Since BIT moved out of the government in 2014 and was rebranded as the 'Behavioural Insights Limited', it has been able to expand its personnel and opened offices in New York, Singapore, Sydney, Toronto and Wellington. Through thousands of workshops and training courses, it has demonstrated

behavioral insights to 20,000 civil servants from all over the world. BIT has also led several international projects on behavioural insights applied for instance on corruption (John 2018, p. 81).

A more fine-grained picture emerges when also considering the actual institutionalisation of nudge units. Different models can be empirically observed (see OECD, 2017; Afif et al., 2018 for an extensive description). Some countries have adopted a steering model with a specialised unit within the centre of government. These units are charged with the task to apply, support or advocate the use of behavioral insights across government. In Germany, for instance, behavioral insights is part of strategic foresight within the Chancellary. The unit works with German government departments to design and implement interventions. Other countries instead opted for a more decentralized specialized model with existing units in departments and agencies focusing on behavioral insights. The UK is probably the most well-known example in this respect, with the dual presence of BIT and departments coordinating their own unit; a clear sign of institutional maturity. A third model is a networked model, as it can be found in the Netherlands and Denmark. In the Dutch case, for instance, each ministry is meant to have its own behavioral insights team, but the Ministry of Economic Affairs has the role of common secretariat. Governments can also merely use behavioral insights for specific projects and initiatives. Finally, but often in addition to the other models, governments may work in partnerships with external institutions. The UK BIT is again a case in point, albeit it is still co-owned by the government. One can see a reflection of a country's political culture and process, in the type of model they opt for, and how they go about institutionalizing it. Yet, countries may change models as time evolves. Interestingly, most countries tend to evolve to a decentralized model, as was also the case for the UK and the US (Afif et al., 2018).

Despite the spread of behavioral insights in public policy, the actual applications remain limited in terms of depth. According to the World Bank review (Afif et al., 2018), most efforts remain limited to an exploratory or pilot phase, or to a particular stage of the policy cycle. Some countries still display risk-averse behavior, as the effects of nudging are not yet widely documented. Also, the application of RCTs which nudges require still constitutes a main challenge for public administrations. Nonetheless, early adopters are clearly maturing and embedding behavioral science via standardized procedures and tools, such as in the UK, or in the Netherlands. As time evolves, it remains to be seen whether we can observe more convergence on this front.

4. Conclusion

While countries have a wide array of evaluation practices and techniques at their disposal, the question is wether we can observe any convergence in terms of the evaluation styles. In this chapter, we focused on three types of practices that are exemplary for regulatory governance and represent different waves of evaluation diffusion. Table 15.3 summarizes our main findings highlighting several dimensions related to the sedimentantion of each evaluation practice associated with different diffusion waves. Although our analysis is by no means representative for all evaluation practices associated with a particular wave, some interesting observations can be made.

Sedimentation of evaluation practice	SCM	RIA	RCT/Nudge
Conceptual clarity	YES, based on an accounting formula	NO	YES, based on concepts and definitions developed in behavioral science
Complete diffusion	YES, but also fading away	YES, but huge variation in the evaluation methods adopted	NO, only few EU countries have adopted nudges
Rapid and coherent implementation	YES	NO	NO
Extensive institutionalization	NO. Mainly because SCM does not require major administrative capacity	Partially. In several countries RIA is still symbolic	NO. With a few exceptions, nudge units are still in the pilot phase
Relevant change of evaluation style	NO	YES, but only in countries that succeeded in the institutionalization	YES, but full impact will become clear in future
Change agents	International networks and EU	International organizations	Pioneer country
Public support	YES	Unclear	YES, although still contested in certain countries

Table 15.3: The extent of sedimentation of regulatory evaluation practices

Altogether, our analysis confirms the relevance of the wave analogy: Countries' evaluation styles are indeed formed by different sediments of evaluation diffusion waves. In determining the extent and size of these sediments, several factors turn out to be important, albeit not necessary per se. First, as the case of SCM shows, conceptual clarity clearly fosters a rapid and uniform adoption across countries. In the case of RIA instead, the lack of conceptual clarity has resulted in a diversity of evaluation methods associated with the approach. Secondly, our analysis confirms the key role of international organisations and networks in disseminating evaluation practices, at least for SCM and RIA. For RCT/Nudge practices, especially pioneering nudge units (i.e., the UK's BIT unit)

took up the role of change agent. Third, public support is a driving factor: as long as governments and citizens of many countries have concerns about its legitimacy, the application of nudging in the public policy process risks to remain at the pilot phase (Afif et al., 2018). Whereas the institutionalisation of evaluation processes takes time (see also Pattyn, 2014), and scholars revealed a negative correlation between years of adoption and institutionalisation, evaluation fashions and sediments also seem to gradually fade away. With NPM having lost momentum, also the popularity of SCM is decreasing. Time will tell which sediments the 'newer' evaluation waves will leave in the longer run.

What do these results tell us with respect of the increased global similarity of policy evaluation styles? What is the actual impact of the adoption of these innovations of regulatory governance and policy evaluation procedures? Our results do not confirm the homogenization expectation of the convergenist camp based on processes of Americanization (Legrand, 2012) or an OECDization of evaluation styles (Magone, 2012; Theodore and Peck, 2012). But they do attest the argument that practices founded on theories and methods—such as RCT/nudge and CBA/RIA—tend to profoundly change the evaluation culture of European public administrations, although the diffusion and institutionalization process is slow and patchy. Conversely, although characterized by a rapid diffusion, the SCM is not founded on a solid theoretical and methodological foundation, which has resulted in limited cultural change and a legitimation issue that undermined its persisting institutionalization. Neither do they disprove the camp of scholars who argue that evaluation systems are still revolving around administrative legal families (Peters, 1997; Peters, 2008). Especially for the actual institutionalization of evaluation practices, our results show that there is evidence of clustered convergence clustered mainly around the capacity of evaluation and auditing institutions. This partial disprove of both comparative camps calls for further theoretically and methodologically assessements of the extent of convergence brought by diffusion waves in regulatory goverance.

References

Afif, Z., Islan, W. W., Calvo-Gonzalez, O., Dalton, A. G. (2018) *Behavioral Science Around the World: Profiles of 10 Countries* (English). eMBeD brief. Washington, D.C. : World Bank Group.

Alkin, M. C. and Christie, C. A. (2004) 'An evaluation theory tree', in M. C. Alkin (ed.) *Evaluation Roots: Tracing Theorists' Views and Influences*, Sage: Thousand Oaks pp. 12-65.

Barbier, J.C. and Hawkins P. (2012) *Evaluation Cultures. Sense-making in Complex Times*, New Brunswick: Transaction Publishers.

Brousselle, A., and Buregeya, J.-M. (2018) 'Theory-based evaluations: Framing the existence of a new theory in evaluation and the rise of the 5th generation', Evaluation, 24(2), 153–168.

Burden, B. C., Canon, D.T., Mayer, K. R. and Moynihan, D. P. (2012) 'The effect of administrative burden on bureaucratic perception of policies: Evidence from election administration", *Public Administration Review*, 72(5): 741–751.

Coletti, P. (2013) *Evidence for Public Policy Design: How to Learn from Best Practice*. Basingstroke: Palgrave Macmillan.

Coletti, P., and Radaelli, C. M. (2013) 'Economic rationales, learning, and regulatory policy instruments', *Public Administration*, 91(4): 1056–1070.

Cowen, T. (2005) Using cost-benefit analysis to review regulation, Paper presented at the New Zealand business roundtable.

De Francesco, F. (2011) 'Diffusion of regulatory impact assessment and standard cost model: a comparative analysis', in L. Mader and M. Tavares de Almeida (eds) *Quality of Legislation: Principles and Instruments*, Baden: Nomos, pp. 238–250.

De Francesco, F. (2012) 'Diffusion of regulatory impact analysis among OECD and EU member states', *Comparative Political Studies* 45 (10): 1277–12305.

De Francesco, F. 2013 *Transnational Policy Innovation: The OECD and the Diffusion of Regulatory Impact Analysis*. Colchester: ECPR press.

De Francesco, F. (2016) 'Diffusion across OECD countries', in C. Dunlop and C. M. Radaelli (eds) *Handbook of Regulatory Impact Assessment*, Cheltenham: Edward Elgard Publishing.

Dimaggio, P. J. and Powellm W. W. (1983) 'The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields', *American Sociological Review*, 48(2): 147-160.

Furubo, J. E. and Sandahl, R. (2002) 'Introduction. A diffusion perspective on global developments in evaluation', in J.E. Furubo, R.C. Rist and R. Sandahl (eds) *International Atlas of Evaluation*, New Jersey: Transaction publishers, pp. 1-23.

Grimmelikhuijsen, S., Jilke, S., Olsen, A. L., and Tummers, L. (2017) 'Behavioral public administration: Combining insights from public administration and psychology' *Public Administration Review*, 77(1): 45-56.

Guba, E. G. and Lincoln, Y. S. (1990) *Fourth Generation Evaluation*. Newbury Park: CA: Sage Publications.

Heidbreder, E., Wegrich, K. and Fazekas, M (2010) 'The double beat of the standard cost model adoption across Europe: How policy diffusion and Europanisation mechanisms interconnect', Paper presented at the third biannual conference of the ECPR Standing Group on Regulatory Governance: Regulation in the Age of Crisis, Dublin.

Jacob, S., Speer, S. and Furubo, J-E. (2015) 'The institutionalization of evaluation matters: Updating the international atlas of evaluation 10 years later', *Evaluation* 21(1): 6–31.

James, O., Jilke, S. R., and Van Ryzin, G. G. (eds) (2017) *Experiments in public management research: Challenges and contributions*, Cambridge: Cambridge University Press.

John, P. (2018). How far to nudge?: assessing behavioural public policy. Edward Elgar Publishing.

John, P. (2019) The international appeal of behavioural public policy: is nudge an Anglo-American phenomenon?, Journal of Chinese Governance, 4:2, 144-162, DOI: 10.1080/23812346.2019.1576264

Lai, M. K. (1991) Field based concerns about fourth-generation evaluation theory. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.

Legrand, T. (2012) 'Overseas and over here: Policy transfer and evidence-based policy-making', *Policy Studies* 33 (4): 329–348.

King, J. A. (2003) 'The challenge of studying evaluation theory', *New Directions for Evaluation*, 97: 57-67.

Majone, G. (1994) 'The rise of the regulatory state in Europe' *West European Politics*, 17(3): 77–101.

Mayne, J. (2012) 'Contribution analysis: Coming of age?', *Evaluation*, *18*(3): 270-280.

Moynihan, P. D., Herd, P. and Harvey, H. (2015) 'Administrative burden: Learning, psychological, and compliance costs in citizen-state interactions', *Journal of Public Administration Research and Theory*, 5(1): 43-69.

Mossberger, K. (2000) *The Politics of Ideas and the Spread of Enterprise Zones*. Washington DC: Georgetown University Press.

Nielsen, M. M., Carvalho, N. R., Veiga, L. G. and Barbosa, L. S. (2017) 'Administrative burden reduction over time: literature review, trends and gap analysis', in *Proceedings of the 10th International Conference on Theory and Practice of Electronic Governance*, pp. 140-148.

Nijsen, A. (2013) 'SCM 2.0: An argument for a tailored implementation', in A. Alemanno, F. den Butter, A. Nijsen and J. Torriti (eds) *Better Business Regulation in a Risk Society*, New York, NY: Springer New York, pp. 231–251.

OECD (1997) *Regulatory Impact Analysis: Best practices in OECD countries*, Paris: OECD publishing.

OECD (2006) *Cutting Red Tape National Strategies for Administrative Simplification*, Paris: OECD Publishing.

OECD (2017) *Behavioural Insights and Public Policy: Lessons from Around the World*, Paris: OECD Publishing.

OECD (2018) *Indicators of Regulatory Policy and Governance*, online <https://qdd.oecd.org/subject.aspx?Subject=GOV_REG>, accessed on 27 January 2020.

Pattyn, V. (2014) 'Why organisations (do not) evaluate? Explaining evaluation activity through the lens of configurational comparative methods' *Evaluation: The International Journal of Theory, Research and Practice*, 20(3): 348–367.

Pattyn, V., Van Voorst, S., Mastenbroek, E. and Dunlop, C. A. (2018) 'Policy evaluation in Europe', in E. Ongaro and S. Van Thiel (eds) *The Palgrave Handbook of Public Administration and Management in Europe*, London: Palgrave Macmillan.

Pattyn, V. (2019) 'Towards appropriate impact evaluation methods' *European Journal of Development Research* 31(2): 174-179.

Pawson, R. and Tilley, N. (1997) *Realistic Evaluation*, London: Sage Publications.

Peters, B. G. (1996) 'Political institutions: old and new', in R. E. Goodin and H. D. Klingemann (eds) *A New Handbook of Political Science* Oxford: Oxford University Press, pp. 205-220.

Peters, B. G. (1997) 'Policy transfers between governments: The case of administrative reforms', *West European Politics* 20(4): 71–88.

Peters, B. G. (2008) 'The Napoleonic tradition', *International Journal of Public Sector Management*, 21(2): 118–132.

Radaelli, C. M. (2005) 'Diffusion without convergence: how political context shapes the adoption of regulatory impact assessment', *Journal of European Public Policy*, 12 (5): 924–943.

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

Radaelli, C. M. and Meuwese, A. C. (2009) 'Better regulation in Europe: between management and regulation', *Public Administration*, 87(3): 639–654.

Rey, L., Brousselle, A., and Dedobbeleer, N. (2012) 'Logic analysis: testing program theory to better evaluate complex interventions' *The Canadian Journal of Program Evaluation*, 26(3): 61-89.

Rossi, P. H. and Freeman, H. E. (1985) *Evaluation: A systematic Approach*, Newbury Park, CA: Sage Publications.

Scriven, M. (1991) *Evaluation Thesaurus*. 4th edition, Newbury Park, CA: Sage Publications.

Shadish, W. R., Cook, T. D., and Leviton, L. C. (1991) *Foundations of Program Evaluation: Theories of Practice*, Newbury Park, CA: Sage Publications.

Stame, N. (2003) 'Evaluation and the policy context: The European experience', *Evaluation Journal of Australiasia*, 3(2): 36-43.

Stockmann, R., Meyer, W. and Taube, L. (2020) *Evaluation Globe*. Palgrave Macmillan.

Sunstein, C. R., Reisch, L. A., and Kaiser, M. (2019) ‚Trusting nudges? Lessons from an international survey' *Journal of European Public Policy*, 26(10): 1417-1443.

Thaler, R. H. and Sunstein, C. R. (2009) *Nudge: Improving Decisions about Health, Wealth, and Happiness*, New York: Penguin Books.

Theodore, N., and Peck, J. (2012) 'Framing neoliberal urbanism: Translating "commonsense" urban policy across the OECD zone', *European Urban and Regional Studies* 19(1): 20–41.

Torriti, J. (2007) 'The standard cost model: when better regulation fights against red-tape', in S. Weatherill (ed.) *Better Regulation*, Oxford: Hart Publishing.

Tummers, L. (2019) 'Public policy and behavior change', *Public Administration Review*, 79(6): 925–930.

Vedung, Evert, 2010, "Four Waves of Evaluation", Evaluation, 16(3), 263–277.

Wegrich, K. (2009) The administrative burden reduction policy boom in Europe. Comparing mechanisms of policy diffusion, CARR Discussion paper No. 52. London: LSE.

Weigel, W. (2008) The standard cost model - A critical appraisal, SSRN Scholarly Paper, online https://papers.ssrn.com/abstract=1295861, accessed on 21 October 2019.

Weiss, C. H. (1993) 'Where politics and evaluation research meet', *American Journal of Evaluation*, 14(1): 93-106.

Widmer, T., and De Rocchi, T. (2012) *Evaluation: Grundlagen, Ansätze und Anwendungen*, Zürich: Rüegger Verlag.

World Bank. 2007. *Review of the Dutch Administrative Burden Reduction Programme*. Washington D.C.: World Bank Group, online <http://www.doingbusiness.org/documents/DBdutch_admin_report.pdf>, accessed on 27 January 2020.