Policy Sci (2018) 51:269–290
https://doi.org/10.1007/s11077-018-9318-4

RESEARCH ARTICLE

Neglected challenges to evidence-based policy-making: the problem of policy accumulation

Christian Adam[1] · **Yves Steinebach**[1] · **Christoph Knill**[1]

Published online: 5 May 2018

Abstract Claims for evidence-based policy-making are motivated by the assumption that if practitioners and scholars want to learn about effective policy design, they also can. This paper argues that this is becoming more and more challenging with the conventional approaches due to the accumulation of national policy portfolios, characterized by (a) a growing number of different policy targets and instruments, that (b) are often interdependent and (c) reformed in an uncontrolled way. These factors undermine our ability to accurately relate outcome changes to individual components within the respective policy mix. Therefore, policy accumulation becomes an additional source of the well-known 'attribution problem' in evaluation research. We argue that policy accumulation poses fundamental challenges to existing approaches of evidence-based policy-making. Moreover, these challenges are very likely to create a trade-off between the need for increasing methodological sophistication on one side, and the decreasing political impact of more fine-grained and conditional findings of evaluation results on the other.

Keywords Policy design · Policy mixes · Policy complexity · Learning · Performance management · Evidence-based policy · Outcome-based learning · Policy evaluation · Policy accumulation

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11077-018-9318-4) contains supplementary material, which is available to authorized users.

✉ Christian Adam
christian.adam@gsi.uni-muenchen.de

Yves Steinebach
yves.steinebach@gsi.uni-muenchen.de

Christoph Knill
christoph.knill@gsi.uni-muenchen.de

[1] LMU Munich, Munich, Germany

Introduction

Policy makers may have a range of different reasons that inform their decisions. Yet, as emphasized by Capano and Lippi (2017), these reasons are typically connected to two main purposes, namely, the search for effectiveness (*instrumentality*) and the construction of a shared sense and common acceptance of a policy (*legitimacy*). It is also well-acknowledged that there might be trade-offs between these two purposes and that the relevance of both instrumentality and legitimacy concerns can vary over time and across policy areas (Adam et al. 2017a). Although instrumentality and legitimacy reflect distinctive rationales guiding policy choices, they both essentially rely upon evidence. On one hand, evidence about the cause–effect relationships is crucial for governments seeking to enhance the effectiveness of their policies. On the other hand, providing evidence for the effectiveness of policy choices is one of the cornerstones of legitimate policy-making. In other words, we all expect policy to be based on at least some evidence, and we tend to assume that if we want to learn about policy effectiveness, we also can.

Yet, it is common sense knowledge that what sounds straightforward in ideal-world scenarios of public policy textbooks can turn out to be a highly demanding exercise in practice. Even if we ignore potential political obstacles to conducting proper evaluations, such as politicians' attempts to selectively use evaluation results for their own purposes (Schlaufer et al. 2018), important methodological challenges do prevail. These challenges stem not only from costly and time-consuming endeavors of data collection, but also from the requirement to develop appropriate research designs that tell us how likely it is that changes in policy outcome are in fact attributable to distinctive choices of policy outputs. These challenges are reinforced by the fact that policy problems are usually highly complex and wicked, with efforts to solve one aspect of a problem potentially creating other one (Rittel and Webber 1973; Alford and Head 2017).

While these difficulties characterizing policy evaluation have long been recognized and been accompanied by a continuous refinement of methodological tools, we argue in this article that an additional challenge for evaluation studies has so far been fairly neglected: the phenomenon of continuous policy accumulation. Policy accumulation comes to light once we stop concentrating on individual policy changes but rather focus on the long-term and aggregate development of policy outputs. Although democracies must balance competing demands and forge compromises, their productivity in accumulating more and more policy instruments addressing more and more policy targets is remarkable. This continuous expansion of the volume of law and regulations has been described in many ways: as an increase in policy density (Knill et al. 2012), policyscapes (Mettler 2016), policy layering (Thelen 1999, 2004), or as an emergence of complex policy mixes (Howlett and Del Rio 2015). Common to all these descriptions is the general trend of continuous policy accumulation as the rate of policy production continues to exceed the rate of policy termination by far (Adam et al. 2017b).

Focusing on policy accumulation highlights new, and so far, neglected dimensions of the well-known 'attribution problem' in evaluation research: First, as policy accumulation makes policy mixes increasingly complex, our job as evaluators does no longer only consist of handling many non-policy parameters as control variables. Rather, the many parameters within the policy mixes need to be handled as well. Thus, the challenge is not only to integrate a relatively high number of control variables, but also to cope with an increasingly complex nature of the key independent variable, i.e., the increasingly complex nature of accumulating policy portfolios. Second, policy accumulation creates an increasing need

for contextualizing policy evaluation. It requires us to evaluate more than just the average effectiveness of a specific policy within a specific policy mix or across different policy mixes. Instead, policy accumulation calls for the evaluation of the conditional effects of individual policy elements. The problem is that this is difficult to accomplish with the conventional methods, and results of conditional effects are very difficult to both interpret and communicate. This makes it very difficult to shape decision makers' thinking about policy issues on the basis of increasingly complex results.

To substantiate our argument, we proceed in three steps. We first summarize the rationale behind and the problems commonly associated with evidence-based policy-making. In a second step, we discuss in which ways interactions of accumulating policy elements within policy mixes contribute to the attribution problem of evaluation research. In so doing, we show the limitations of existing approaches to theoretically or methodologically resolve the challenges posed by policy accumulation. We conclude with a discussion of the general implications of our findings for the prospects of evidence-based policy-making. We argue that there is a potential trade-off between the achievement of more fine-grained evidence via methodological sophistication on one side, and the decreasing political impact of such findings in day-to-day policy-making on the other.

Empirically, we focus on the areas of environmental and social policy in 21 OECD countries in the period between 1975 and 2005. While we concentrate on these two sectors, we believe that the presented '*independent variable problem*' caused by constant policy accumulation is in fact of general relevance. To emphasize this general relevance of our argument, we include examples from other policy sectors wherever plausible. Finally, we close with several reflections on the implications of our argument.

In addition to the study of policy evaluation and evidence-based policy-making (e.g., Anderson and Rees 2013; Pacula et al. 2015), our paper contributes to the literature on policy design (e.g., Howlett and Lejano 2013; May 1991; Schneider 2012). Recent calls for a renewal of this literature asked for the development of "greater conceptual clarity and the methodological sophistication needed to sift through the complexity of new policy regimes" (Howlett and Lejano 2013, 369). Furthermore, this renewal should help to remind ourselves that while certain policy instruments are often discussed as panaceas, their effects strongly depend on the context in which they are employed (Howlett and Lejano 2013, 370). In this article, we address both issues. On one hand, we capture the increasing size and complexity of social and environmental policy portfolios in a systematic way that can be used for other policy sectors as well. On the other hand, we discuss how the observed complexity makes it more and more difficult to account for all (potentially) relevant policy parameters and parameter interactions with conventional tools.

Increasing policy effectiveness through evidence-based policy-making?

The major goal of policy evaluation is to analyze whether a certain policy achieved its intended effects and to learn how the policy design could be improved and optimized to further increase policy effectiveness; i.e., the extent to which a given policy achieves its intended objectives (*policy outcomes*). Evaluations can serve as a basis for deciding on the continuation or termination of a policy and its substitution or amendment (Rossi et al. 2003; Bauer and Knill 2012). Evaluation is thus essential to any form of evidence-based decision-making, building on the premise that "[d]ecision-makers will learn from performance information, and, in turn, they will make better-informed decisions and

improve government performance" (Moynihan 2005, 203). Policy evaluation thus constitutes a central element that is used to help politicians make well-informed decisions about policies by putting the best available evidence at the heart of policy formulation and implementation (Howlett et al. 2009; Head 2016). In other words, these approaches assume that if we want to learn about effective policy design, we also can.

The systematic reliance on policy evaluation and evidence-based policy-making has been promoted over the last three decades, well in line with the spread of new ideas of how to organize public-sector organizations, such as the new public management (NPM) approach. One of the key elements associated with these ideas was to shift the way that policy performance is measured from processes to results (Radin 2009).

In general, results can be measured on at least two different levels: outputs and outcomes. Outputs are products, goods, and services provided by public-sector organizations on the basis of public policies, such as testing water quality, paying out child benefits, conducting traffic controls, and providing measles vaccines. In contrast, outcomes refer to more general societal or environmental developments that are assumed to be, at least in part, the result of policy outputs. They correspond to the general policy goals to be achieved. Outcomes corresponding to the exemplary outputs listed above could be the level of biodiversity in rivers and lakes, the number of newborn children, poverty rates among young families, the number of alcohol-related traffic accidents, and the number of measles-related deaths.

While outputs as indicator bear the advantage that they can be directly attributed to organizational activity, outputs do remain means, not ends. They remain means that have the potential to contribute to the achievement of broader policy goals, but this potential is not necessarily always exploited. Therefore, focusing on outcomes promises to capture a more crucial dimension of results: the level of actual goal attainment or policy effectiveness. Focusing on effectiveness is hence perceived as a way to inform more intelligent and effective policy designs (Davies 2012; Glennerster 2012). The assumption is that evaluating whether and how specific policies are able to achieve their aspired goals ultimately helps to optimize policy arrangements and to develop better policies through the promotion of evidence-based policy-making.

In addition, indeed, approaches of evidence-based policy-making have gained some prominence in several countries over the last decades, although the extent to which these concepts are applied varies strongly across countries (Barnow 2000; Favero et al. 2016; Heinrich 2002; McBeath and Meezan 2010; Radin 2000). In particular, in Anglo-Saxon countries, such as the United Kingdom or the United States, the incorporation of evidence into the policy formulation process is common, and detailed guidelines have been produced to standardize their use. The Government Performance and Results Act from 1993 represents just one of many earlier and later initiatives of the US government intended to strengthen the focus of policy makers and implementers on results. Similarly, the United Kingdom, for example, introduced Public Service Agreements (PSA) in 1998 to implement outcome-based performance schemes. In these agreements, central government departments (and hence many of the bodies they fund) commit themselves to reach specific outcome goals. In 2004, the Swiss government adopted an overall strategy for the further development and expansion of FLAG ('Führen mit Leistungsauftrag und Globalbudget') at both the central and regional levels. In particular, regional governments like the canton Zug use performance agreements ('Leistungsaufträge') that define, for example, environmental outcome targets, such as the goal to increase the share of construction waste suitable for recycling or to lower phosphate levels in Lake Zug to a certain degree.

While approaches of evidence-based policy-making have been strongly promoted by scholars of public policy and public management, they have not remained without criticism (Van Thiel and Leeuw 2002). In this context, we can broadly distinguish between three types of criticism: one focusing on *measurement*, the other one focusing on *implementation*, and one focusing on *the attribution of causes*. The first critique holds that outcome measures are often poorly designed or chosen and, therefore, fail to capture relevant outcomes in an adequate way (Radin 2006; Piotrowski and Rosenbloom 2002). The second critique proposes that even when relevant measures are found, they are often imperfectly integrated into decision structures and processes relevant for policy implementation and policy-making (Van Dooren and Van de Walle 2016; Mayne 2007). Both these critiques still imply that with sound efforts to create appropriate measures and efforts to use the collected information in decision-making processes, outcome measurement can help us to learn about effective policy design. The third critique questions this assumption (Bovaird 2012; Pollitt 2011). Specifically, it states that even when we have valid measures and proper implementation, we are still confronted with the problem of attributing outcome changes to causes: are the outcome changes caused by public policy? Which elements of the policy portfolio contributed by how much to the observed outcome changes?

It is generally considered that the source of this 'attribution problem' is that we typically operate outside of experimental settings with controlled environments and randomized stimuli. Instead, we tend to often rely on observational research designs and naturally occurring policy experiments. This implies that we need to control for all influences on the policy outcome other than just public policy. The conceptual and methodological toolkits for designing and evaluating public policy all reflect this attempt to isolate policy effects by controlling for as many other influences as possible. Thereby, they help to alleviate the 'attribution problem'. Conceptual tools for policy design, such as strategy maps, for example (Kaplan and Norton 2004), reflect visualized hypotheses about cause-and-effect relationships that ideally try to include all factors affecting the relevant outcome. Similarly, the methodological toolkit for policy evaluation—including comparative case study designs, regression discontinuity methods, difference-in-differences methods as well as panel methods—all attempt to isolate the actual impact of public policy by controlling for potentially confounding influences.

While our article joins this choir of authors emphasizing problems of causal attribution, our main objective is to highlight a different source of this 'attribution problem': as policy accumulation makes policy mixes increasingly complex, our job as evaluators does no longer consist only of handling many non-policy parameters as control variables; the many parameters within policy mixes have to be handled adequately as well. Thus, the challenge is not only to integrate a relatively high number of control variables, but also to cope with an increasingly complex nature of the independent variable—the increasingly complex nature of public policy.

Evaluating complex policy mixes: potential solutions to the 'independent variable problem' and their limitations

Our argumentation starts from the observation that policy spaces are rarely empty. Rather, new policies in a certain sector continuously add up to an already existing mix of policy targets and policy instruments. While this insight in some ways reflects a truism in the public policy literature, the consequences of this development of policy accumulation for

evaluation and evidence-based policy-making have not been systematically addressed so far.

Policy accumulation occurs whenever policy makers adopt new rules without abolishing others. If law makers decide to replace one policy element with another, the size of the policy portfolio remains stable. Likewise, if states drop a policy element without introducing a replacement, a policy portfolio reduces in size. Accordingly, our notion of policy accumulation is strongly connected to the concept of policy layering (Schickler 2001; Thelen 1999), i.e., a process of institutional evolution in which institutional arrangements are gradually enhanced with new elements, while the pre-existing institutional structure, which has become entrenched through the vested interests defending it, remains stable. In line with Thelen (2004), we argue that such processes of policy layering are pervasive in modern democracies, and while the motivations that drive policy layering are manifold, policy layering inevitably leads to policy accumulation.

We conceive of policy accumulation as continuous addition of new policy elements to existing policy portfolios without the compensatory reduction of already existing ones. In this context, a policy element constitutes the combination of a policy target and a policy instrument. While policy targets define what or who is being addressed by a new policy, policy instruments define how the target is being addressed (Eliadis et al. 2005). To understand the challenges of evidence-based policy-making created by policy accumulation, we have to answer the question of how much the individual elements within policy mixes account—independently or in combination with other elements—for the observed level of outcome achievement. On this basis, we can decide which elements of the relevant policy portfolio to dismantle, which elements to keep, and which new elements to integrate.

Quite intuitively, we argue that as the number of elements in the policy portfolio increases, it becomes harder to disentangle whether and by how much these different elements contribute to any observed outcome change. This is because we not only have to take a greater number of individual policy elements into account. Typically, we also have to consider the potential importance of interactions between these individual elements: we have to consider more policy parameters and parameter interactions. As a consequence, it is harder to draw correct conclusions about the contribution of individual parameters to outcome change. This is because of the way in which we typically try to make such inferences: policy learning becomes more difficult because of the conventional ways in which such learning occurs. These conventional approaches typically are ex-ante cogitation and the ex-post evaluation based on observational research designs. While these approaches can be of enormous help in many contexts, they tend to become more and more problematic when they have to cope with an increasing number of parameters.

As will be shown in the following section, policy accumulation creates problems of causal attribution that so far have not only been fairly neglected, but are also difficult to handle. To underscore the distinct character of this argument, we refer to such accumulation-induced problems of causal attribution as the ‘independent variable problem’. It is important to understand that the presented ‘independent variable problem’ is different from the common attribution problem and that the main issue rests with the extent to which individual policy elements can really be seen as independent from one another.

First, the independent variable problem goes beyond problems of sufficiently taking account of control variables, which is commonly known as the ‘attribution problem’. Technically, control and independent variables are of course all explanatory variables. However, control variables denote variables that are taken into account merely to identify the effects of the theoretically interesting independent variables. In contrast, we describe the independent variable problem as one of handling the increasing complexity of the independent

variables, that is, of managing the accumulating array of interacting policy targets and instruments in existing policy portfolios.

Second, the term 'independent variable problem' highlights that it is not only the number of independent variables—the number of policy targets and instruments—that makes them hard to handle. It is also the assumed independence of these policy targets and instruments that is problematic. Our independent variables (the components of public policy) are thus often not independent—neither in their occurrence nor in their effects. Adding new or abolishing existing elements of the policy mix takes place under the consideration of the other parts of this policy mix. Furthermore, policy elements do not exert their influence independent of each other, but their effects are often conditional on the configuration of other policy elements. The foundation of evidence-based policy-making in this situation becomes the identification and understanding of conditional policy effects as opposed to average policy effects. As policies accumulate, the proper assessment of these complex relations tends to challenge both our cognitive and methodological capacities.

If we accept the basic argument that policy accumulation poses strong (and potentially growing) challenges for evidence-based policy-making, the immediate question that arises is how such challenges might be addressed. Indeed, we can identify several approaches that are applied to cope with accumulation-induced challenges, including (1) theoretical ex-ante assessments of interactions; (2) modeling of interactions; and (3) attempts of keeping interaction factors stable. Yet, each of these approaches can be considered merely as a partial remedy to address the challenges emerging from policy accumulation.

'Careful examination' of increasing policy portfolios?

While the relevance of interactions between different elements within the policy mix is well known to the literature on policy design and policy mixes, this literature tends to assume that the effects of policy mixes can be assessed ex-ante on theoretical or conceptual grounds (Linder and Peters 1989; McConnell 2010; Del Río 2014). For example, where policy mixes were found to be coherent, consistent, or congruent, they were assumed to have the intended impacts (Howlett and Rayner 2013). The key assumption underlying this approach is that with 'careful examination' (Howlett and Lejano 2013, 362), appropriate cogitation, and instrumental learning (May 1991), we would be able to 'improve policy designs and outcomes' (Howlett and Lejano 2013, 362).

To demonstrate that such increases in the number of relevant policy elements over time are quite substantive, we take a closer look at accumulation rates of environmental and social policy portfolios. We chose these policy fields as they constitute major areas of state intervention in modern democracies and allow us to compare complexity dynamics in two fields that had reached very different degrees of maturity at the beginning of our observation period. While social policy represents a well-established field, environmental policy only emerged as a distinct field in the 1960s and 1970s. In these areas, we use an original data set covering the environmental and social policy outputs of 21 OECD countries over a period of three decades (1976–2005). The data were compiled under the CONSENSUS project financed under the European Commission's 7th Framework Programm, with the help of national experts and then coded by the project members to guarantee high quality, validity, and reliability. To make sure that policy changes were classified the same way, different intercoder reliability tests were performed. As far as the exact codification is concerned, a detailed coding manual helped to extract the relevant information from the legal documents in a systematic and comprehensive manner. The policy change assessment relies on a comprehensive data collection encompassing all relevant national legal

documents—laws, decrees, and regulations—in the specific policy areas under review. The legislation was collected through national legal repositories, secondary literature, and scholarly analyses. Legislation emanating from subnational level was excluded from the data collection process.

Even with an extensive investment of resources, we were unable to assess both policy fields in their entirety. We thus had to limit our analysis to the subfields of clean air policy, nature conservation, and water protection for environmental policy, and to the subfields of children, pensions, and unemployment for social policy. Across these subfields, we distinguish between 48 environmental policy targets that can potentially be regulated. These include, for example, regulation of ozone, carbon dioxide, or sulfur dioxide in the air. Furthermore, environmental policy targets include the regulation of lead content in gasoline, or also the sulfur content in diesel as well as nitrates and phosphates in the continental surface water, or regulations regarding native forests, endangered plants, and endangered species. These targets can be addressed by none, one, or by several different instruments. We distinguish between 12 instruments, such as obligatory standards, the setting of technological prescriptions, the introduction of a permit, or the payment of a subsidy.

Turning to the field of social policy, policy targets range from the basic unemployment of individuals, over pensions for individuals and also for married couples, all the way to children and juveniles. In total, we examine 19 social policy targets. Each of these policy targets can be addressed by one or by several policy instruments. Again, these targets can be addressed by none, one, or by several instruments. We distinguish six social policy instruments including contributions to be paid, one-time bonuses, regular allowances, or tax exemptions, for example. A full list of all environmental and social policy targets as well as a list of the 12 environmental policy instruments and the six social policy instruments is provided in the online appendix (A.1–A.4). Of course, the number of relevant policy targets and instruments that can potentially be addressed in legislation varies across different policy areas. Therefore, we abstain from absolute comparisons of policy portfolio size between these two sectors.

Figures 1 and 2 present the number of policy targets and instruments in place over time for the area of environmental and social policy across our country sample. These figures illustrate two developments. First of all, the number of policy targets addressed within both policies increased in almost all countries under study. In the area of environmental policy, the average number of policy targets increased from roughly seven in 1976 to more than 30 in 2005. While this upward trend is clearly not as pronounced in the area of social policy—a fact that should not be surprising given the much higher degree of maturity of this policy field compared to environmental policy—the average number of social policy targets addressed also climbed from 7 to 12 during our investigation period. Second, the figures show across both sectors and all countries that the number of applied instruments grew far stronger than the number of targets. The average number of environmental policy instruments grew from roughly eight in 1976 to almost 80 in 2005. For social policy, the average number of policy instruments in place increased from 10 to 25 over our investigation period. This clearly reflects the fact that policy makers have started to address existing as well as new policy targets with more than only one policy instrument. More specifically, the ratio of policy instruments and policy targets increased from roughly 1.4 instruments per target in the average country in 1976 to about 2.1 instruments per target in 2005 for social policy. Similarly, the instrument target ratio increased from about 1.14 instruments per target to 2.6 for environmental policy. While we summarize these trends for the relatively broad categories of environmental and social policy portfolios, the tendency to increase the number of policy targets and instruments is by no means restricted to

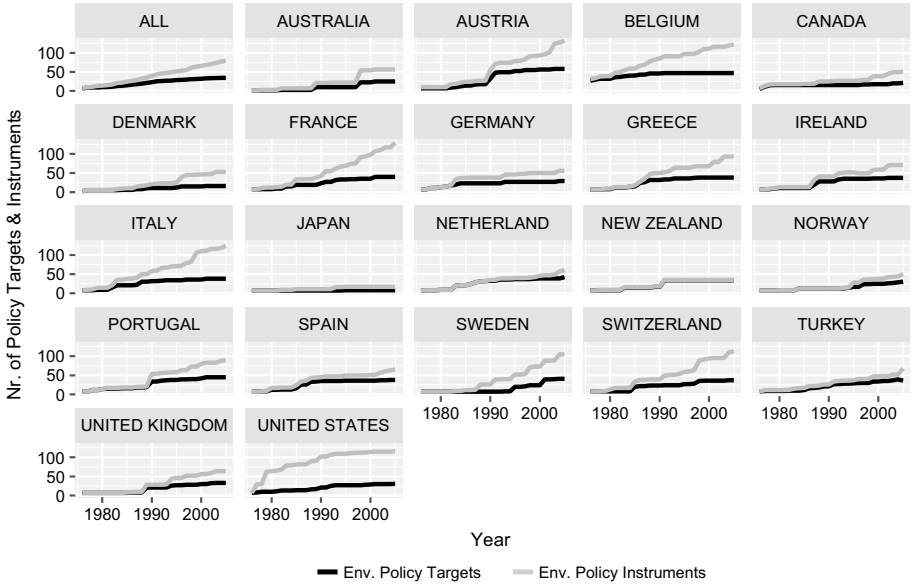

Fig. 1 Number of environmental policy targets and instruments over time (1976–2005)

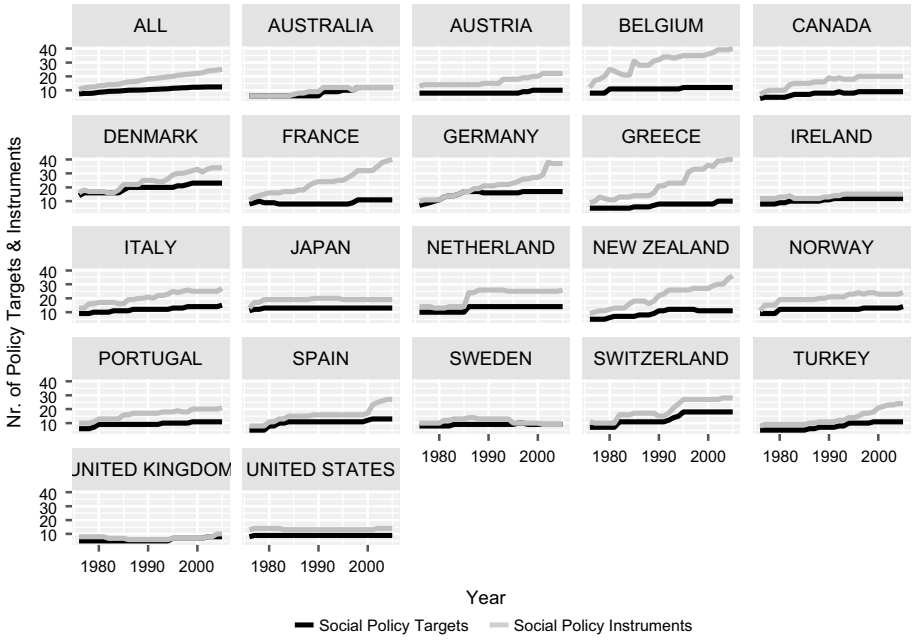

Fig. 2 Number of social policy targets and instruments over time (1976–2005)

any particular subfield within each sector. Instead, portfolio growth is a widespread feature within and across portfolios; however, wide or narrow their boarders are drawn.

Overall, our data reveal a consistent picture of strong increases in policy portfolio size reflected by an increasing number of both policy targets and policy instruments, as well as by an increasing number of instruments per target. This increase can be observed across almost all countries under observation with very few exceptions. Knill et al. (2012) have referred to this increasing policy portfolio size as increasing policy density as more and more policy targets and instruments populate a specific policy sector.

We argue that two consequences emerge from these patterns. First, the disproportionate increase of policy instruments makes any assessment of instrument choice and related outcomes more complicated. We need more time, resources, and knowledge to identify all measures related to the outcome of interest. Omitting some of these measures creates biased results. Second, increasing portfolio size does not only make domestic policy approaches more complicated, but it also makes them more complex. In systems' theory, systems are perceived as complex when 'perfectly understanding the behavior of each component part of a system [will not help to] understand the system as a whole' (Miller and Page 2009, 3). We suppose policy portfolios to be complex in this sense of the term, as the effect of different policy targets and instruments does not follow a simple additive logic. On the contrary, policy targets and instruments are interrelated, and their effects are interdependent. Effects are thus conditional and nonlinear. These interactions, which turn large and complicated policy portfolios into complex ones, fundamentally question our theoretical and conceptual tools for ex-ante cogitation. This is in particular the case as even the same policy mixes might be interpreted and treated in a variety of ways across different institutional setups. As shown by Moulton and Sandfort (2017), the ultimate 'meaning' of policy targets and instruments essentially evolves from the dynamic interactions between multiple social actors being located at different 'implementation levels', i.e., the policy field, the organization, and the frontline.

Explicit modeling of portfolio interactions?

According to Majone, interaction effects within policy portfolios are almost inevitable whenever we are confronted with 'policy space[s]', which are of limited size and a 'population of policies that grows relative to the size of the space' (Majone 1989, 169). In these settings, it becomes less and less avoidable that the consequences of one policy interfere with the operation of another. From this perspective, policy interaction is a result of 'congestion' (Majone 1989, 159) due to increasing portfolio size or policy (Knill et al. 2012). While Majone wrote about interaction as a phenomenon that was inevitable and undesirable, more recent research has also discussed the fruitful and beneficial co-existence of policy instruments (Howlett and Rayner 2013; Neil Gunningham and Grabosky 1998). We consider both types of interactions to be relevant and to co-exist. While some interactions will lead to an unintended weakening of individual policy components' effects, other components will reinforce each other's effectiveness. If we want to identify the accurate effect size of individual components within the policy mix, we need to account for both types of such interaction effects accurately. Specifically, we have to account for potentially relevant interactions (a) between instruments addressing the same policy target as well as (b) interactions between policy instruments addressing different policy targets but the same outcome.

An alternative approach to ex-ante cogitation of interaction effects is hence to explicitly model such interactions within policy mixes and thereby bring these factors to the forefront. By treating policy mixes as variables whose interactions are explicitly considered within statistical models of outcome changes, the independent variable problem of evaluating policy mixes might be reduced. While this might very well be true for policy mixes of low and moderate sizes, the following discussion will illustrate why this strategy soon becomes unviable in the light of more complex policy mixes.

Figure 3 shows what happens under increasing policy complexity. It reveals that even incremental changes in the policy portfolio lead to the exponential growth in the number of parameters to consider. This is because the new components have to be taken into account individually at the same time that we have to assess their (potential) interaction with pre-existing elements of the policy mix. After all, the possibility of beneficial and detrimental policy interactions is the premise on which the whole literature on policy mix and policy designs builds (e.g., Blonz et al. 2008).

To illustrate the relevance of policy interactions and increasing policy portfolios, consider a simple example from a field in which the outcome effect of public policy has come to be heatedly debated in recent years, namely, the regulation of cannabis (Anderson and Rees 2013; Pacula et al. 2015; Adam and Raschzok 2017). Without trying to enter the substance of this cannabis debate, one of the key questions in this context is whether and how strongly the prevalence of cannabis abuse (outcome) depends on whether the recreational consumption of cannabis is banned.

If there were just this one policy target addressed by just this one policy instrument within this mix, no interaction would have to be taken into account. With two instruments (A, B) addressing this policy target, one interaction—the one between them—might already be relevant and would have to be taken into account in addition to the two main effects. For instance, several governments implement different rules for the possession and the consumption of cannabis. With three instruments (A, B, C), four potentially relevant interaction effects have to be integrated, since both bipartite combinations (AB, BC, AC) and multipartite combinations (ABC) are possible (see Fig. 3). For example, states may pursue different rules for the possession and consumption while also sanctioning violations in highly diverse ways. This exponential trend is further enhanced when we consider that

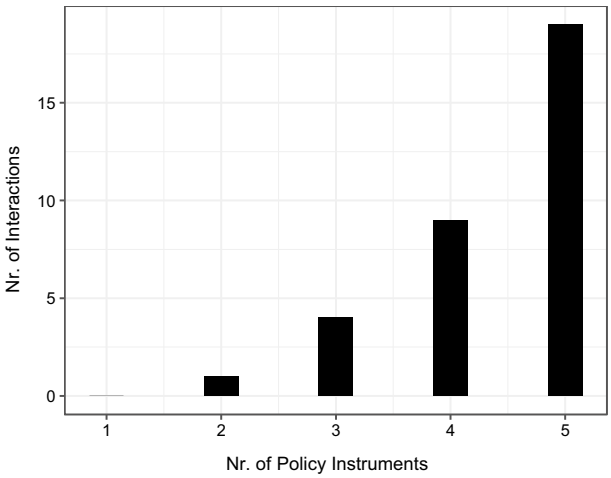

Fig. 3 Number of potentially relevant interactions

interaction effects are not necessarily restricted to instruments addressing the same target. Often, the interaction of several different targets affects the same outcome as well.

In our brief cannabis-related example, some might still feel comfortable trying to disentangle how policy components affect the observed outcome individually and combined through mere reflection. Certainly, however, the addition of further components to this policy mix, such as large-scale information campaigns about the potential dangers of cannabis abuse, will diminish rather than enhance this perceived level of comfort. Conceptual tools for policy design, such as strategy maps (Kaplan and Norton 2004) that try to include all factors affecting the relevant outcome, become more and more difficult to handle effectively as the number of relevant factors increases. Even for large-*N* statistical approaches, identifying actual policy effects becomes more problematic as policy complexity increases. In our rather simple illustration regarding cannabis policy, the statistical representation of the assumed effect structure would already require the estimation of coefficients for seven parameters without even considering any additional non-policy control variables from the socio-economic environment. This is because the statistical representation would have to take into account not only the main interaction but also all constituent elements of the interaction term (Brambor et al. 2006). In this context, this amounts to the four interaction terms plus the three individual policy components.

In her well-cited empirical evaluation of the effectiveness of US state energy programs, Carley (2009) does openly acknowledge the analytical and methodological challenges posed by policy accumulation. While the author considers the influence of "other energy policies" (p. 3073) and "policy interactions" (p. 3076) on the effectiveness of renewable portfolio standards in the analytical framework, the article's empirical analysis is restricted to the analysis of main effects. The author justifies this decision by arguing that the included "variables are [only] crude representations of supporting policy instruments" but that only such method "allows (…) to include these variables in the model without compromising degrees of freedom or unnecessarily complicating the variance–covariance matrix used for model estimation" (p. 3077). In a similar vein, Bassanini and Duval (2009) accept the need to consider both the interaction between each individual policy (AB, BC, AC) and the overall reform complementarities (ABC) to accurately evaluate the effect of social policies on unemployment rates. Yet, in their empirical analysis, they actually opt for two sets of models: one for assessing the "simple interactions" (ibid: 48) between different social policies and one for evaluating the "systematic interactions" between social policies and the "sum of [all] direct effects" (ibid: 51). Again, the selected approach is primarily justified by the problem of handling too many parameters.

While the number of parameters that must be included in the model might still be completely unproblematic in some contexts, we have to consider that we mostly have to operate with relatively few observations when evaluating outcome effects. This applies in particular to studies using time series cross-sectional data with observations representing state- or country years. In these settings, the over-fitting of regression models becomes a real threat when too many parameters are included (Babyak 2004; Goggin 1986). Simply put, over-fitting means that we are asking too much from the available data. In other words, "[g]iven a certain number of observations in a data set, there is an upper limit to the complexity of the model that can be derived with any acceptable degree of uncertainty. Complexity arises as a function of the number of degrees of freedom expended (the number of predictors including complex terms such as interactions and nonlinear terms) against the same data set during any stage of the data analysis" (Babyak 2004, 411). The combination of small sample sizes and many independent variables in the form of policy parameters makes it highly unlikely to identify true effects, let alone true effect sizes for the individual components of

the policy mix. It is important to highlight that this argument is in fact independent from the exact data structure. Even when using a multi-level or nested data structure, one might gain more observations at the individual level, but the restricted number of macrolevel units still prevents testing for the various possible combinations of different policy instruments in a rigorous manner (Möhring 2012).

This discussion is not new, of course. It directly relates to on-going debates in the literature on policy design and policy mixes (Linder and Peters 1989; Neil Gunningham and Sinclair 1999). This literature was built on the insight that policy effects would often depend on the interplay between different elements in the domestic policy mix and not only on the selection or aggregation of individual instruments. Furthermore, this literature has well recognized that policy mixes tend to become increasingly complex (Howlett and del Rio 2015). Our argument, however, goes beyond these rationales provided in the policy design literature. We argue that conventional approaches to evaluate the effectiveness of policy mixes are less and less able to handle the increasing complexity of policy mixes; particularly when the goal is to determine how effective individual elements within these mixes are.

While we consciously chose our example to illustrate policy interactions most plausibly, there would be no design and performance assessment problems—and no need for this paper—if interactions would always be of the nature presented above: plausible, limited in number, and anticipatable. In light of increasing complexity, scholars are increasingly likely to reach their cognitive and methodological limits in handling these interactions: over 30 targets addressed by 80 policy instruments addressed in the average environmental policy portfolio is a lot to sort through. Of course, not all interactions between these elements will be relevant. Yet, separating the relevant from the irrelevant interactions becomes increasingly challenging if not impossible.

Treating policy mixes as stable background factors?

Because distinct policy elements interact within increasingly complex policy mixes, policy accumulation can make it extremely difficult to attribute outcome changes to these individual elements. Neglecting this context inherently undermines the internal validity of evaluation studies by introducing omitted variable bias.

Yet, there are of course-sophisticated approaches that are able to avoid this bias. Quasi-experimental approaches to policy evaluation perform particularly well in ensuring the internal validity of findings. These approaches treat policy mixes as background factors that are kept constant across research subjects. This works especially well in studies that use panel data on individual behavior before and after specific policy reforms of interest. With such approaches, the behavior of individuals affected by the policy can be compared to a quasi-control group of unaffected individuals. Furthermore, the complexity of the policy mix within which the policy reform took place is identical for both groups. This ensures the internal validity of claims about the effect of being exposed to a certain policy. Even as the continuous process of policy accumulation makes policy mixes increasingly complex, this research design should protect the internal validity of findings.

Unfortunately, the strategy of keeping background factors stable comes along with several limitations. First of all, this strategy only helps to address problems of internal validity of outcome-based approaches, but (1) fails to address problems of external validity. Two additional aspects further complicate the research strategy outlined above: (2) the simultaneous treatment of whole populations and (3) the simultaneous addition and change of many aspects within policy mixes.

Problems of external validity

Keeping interactions between elements of policy mixes stable across both groups mitigates problems with internal validity only, not with the external validity of claims. External validity refers to the generalizability of claims and is thus a crucial prerequisite for policy transfer between countries and jurisdictions. Simply put, showing that a specific policy was effective within the context of one particular policy mix hardly allows for predictions about how this policy will perform within a different policy mix. Accordingly, while the identification of policy effects is possible in such research designs, we should be careful in extrapolating these findings to other contexts. External validity problems of policy-based evaluations are thus closely linked to problems of policy transfer (Dolowitz David and Marsh 2002), as is illustrated by the following examples.

In 2009, the European Union adopted the Euro IV norm re-setting the obligatory standards for emissions of heavy-duty vehicles. For the approval of new vehicles, the new emission limits entered into force in 2013. Since 2014, they apply to all registrations. The primary goal of this measure is to substantially reduce exhaust gas emission of carbon monoxide by the road haulage industry (outcome). The speed at which old and thus heavier polluting trucks have been replaced by new ones varies across Europe. This is because of the way the newly introduced instrument (emission standards) interacts with established instruments addressing road haulage emissions. Some but not all member states complemented the introduction of the Euro IV norm with economic instruments offering subsidy payments. Buyers of new trucks received a government subsidy lowering the effective purchase price of a new vehicle complying with the regulation. Moreover, countries such as Germany and Austria reduce levies in the form of road tolls for trucks that meet Euro IV. Thereby, these countries set a further financial incentive to replace older transport fleets and to meet new standards more rapidly (Albalate et al. 2009). We should thus not be surprised that the adoption of the Euro IV norm is associated with different outcomes in different countries. These various effects do not represent random variation but depend systematically on the other elements within the domestic policy mix. Simply calculating the average effect of this norm across all member states without taking its interaction with other policy components into account would be misleading. Average effects result from overestimating effectiveness in countries like Austria and Germany and underestimating its effectiveness in other countries. As long as we do not explicitly account for the conditional nature of the Euro IV norms effect, we will obtain biased results.

Such conditional policy effects are not restricted to environmental policy. They also play an important role in social policy. In terms of outcomes, programs like the child support enforcement program in the US are intended to, inter alia, positively affect the poverty status of recipient families. At first glance, one might assume that a program leading to a better enforcement of child support payments does ultimately also reduce families' poverty status. Yet, this effect might not be as unconditional after all. Pirog and Ziol-Guest (2006) highlight that most scholars failed to consider "the possibility that child support receipt may alter support recipients' labor market participation, such that the total effect on poverty may be either smaller or larger than found in the studies" (p. 970). The decision of labor market participation in turn is likely to depend on the general generosity of welfare benefits and unemployment benefits, on the conditionality of these benefits on receiving child support, on whether labor market placement is more or less protective of single parent households, and on whether services such the provision of kindergarten are free, cheap, or expensive. In case that other welfare benefits are paid independent of household income

through child support, with highly protective policies that do not sanction single parents when refusing to take available jobs, and free kindergarten facilities, enforcing child support payments might in fact increase household income sufficiently as to reduce labor market uptake. In contrast, where enforcing child support payments are coupled with lowering welfare eligibility, expensive and obligatory kindergarten, and substantial sanctions when available jobs are turned down, enforcing child support payments might have actually the opposite effect on labor market uptake.

Which of the presented scenarios will ultimately improve families' poverty status depends on the relation between obligatory child support payments and the average wages paid for available jobs. Against this background, let us assume that we observed an improving poverty status of recipient families after the child support enforcement program was installed. Would we be able to learn from this experience that this program is an effective policy that would have similar effects in other jurisdictions? Obviously, this would only be possible if we estimated the 'correct' effect size while taking all relevant interaction effects into account.

Problems of reform scope

Public policy reforms often affect the whole population simultaneously, which makes it difficult to form a control group. Simply put, when everyone is simultaneously affected by the policy reform, they can no longer be separated into distinct treatment and control groups. Consider pension reforms, health care reforms, or education system reforms. For these 'treatments', it is often hard to separate citizens according to whether or not they are exposed to the reform, at least not in a meaningful way. To learn about the effects of high school reforms, it makes little sense to compare the behavior of high school students with the behavior of retired people (even though the latter are not really affected by the reform). Instead, it is necessary to compare a group of students affected by the reform with another group of students unaffected by it.

One promising way of addressing this challenge has been experimental roll-outs of public policies in the field. This strategy explicitly builds on controlling who is affected by the reform and who is not. While this approach is rightfully considered to be the gold standard in terms of internal validity, it suffers from the same challenges with respect to external validity as quasi-experimental approaches. Whether and how reforms will perform within different policy mixes remains uncertain, and policy accumulation makes the ex-ante anticipation thereof increasingly difficult.

Problems of reform size

Furthermore, policy accumulation is the result of public policy reforms that typically come in the form of 'mixed treatments'. This means that reforms typically entail several simultaneous additions and changes to the existing policy mix. One might think that one factor that could alleviate the problems induced by increasingly complex policy portfolios would be to reform these portfolios in a more controlled manner in the sense of only changing single elements in the policy portfolio.

By changing only one element of the policy portfolio at once, one could approximate conditions of a controlled quasi-experiment and rely on pre–post comparisons to attribute changes in outcome achievement to the reformed element. Unfortunately, however, as our data reveal, policy change does in reality hardly follow this logic. Figure 4 shows—for the

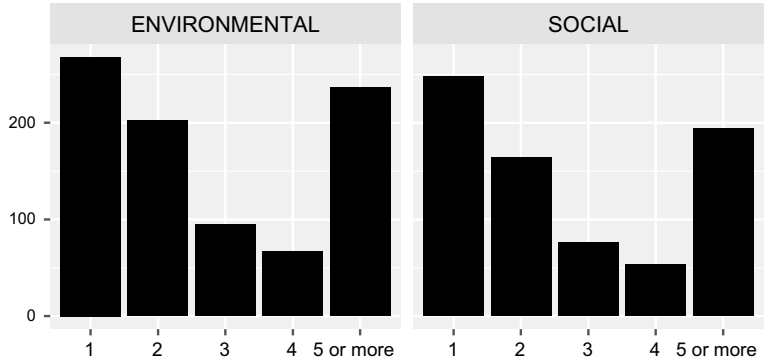

Fig. 4 Frequency distribution of social and environmental policy changes

areas of environmental and social policy—that even when governments adopt rather incremental reforms, they often change more than only one element of the policy mix at a time. Given the contested space on the parliamentary agenda and the restricted time in office, changing individual aspects of the policy mix to be able to assess the effects of this change thus simply does not seem to be a feasible political strategy.

Figure 4 presents the frequency distribution of the annual number of policy targets and instruments reformed simultaneously in all countries and policy subfields under scrutiny. These presented changes include both the introduction and abolishment of policy targets and instruments as well as the recalibration of instrument settings. It becomes apparent that reforms changing only individual policy elements account for not more than about one-third of all reforms observed across the two policy areas. Similar to our arguments presented above, the complexity of policy reforms tends to increase with the number of policy elements changed and the relevance of interactions of these changes with other changes as well as with established, i.e., unreformed policy elements.

Consider, for instance, the German 'Agenda 2010' reform package adopted in 2003. This reform (1) established a means-tested benefit system for long-term employment, (2) reformed the German labor agency, (3) introduced active labor market policies, (4) altered the provisions for laying off employees, and (5) lowered social security contributions for jobs with marginal income (Horn and Logeay 2010). While it has come to be rather undisputed that the policy reforms contributed somewhat to Germany's economic recovery (Rinne and Zimmermann 2012), it is difficult to disentangle which reform elements are important and which ones might have even been counterproductive. Due to the numerous instruments changed, it is hardly possible to isolate the effects of specific components of the reform package on outcome measures such as unemployment rates, wage levels, job matching productivity, etc. Even more so, if the possibility is taken into account that the effects of these reformed elements are conditioned by their interaction with established policy elements in Germany, one will probably become skeptical of whether similar changes will also yield comparable effects in other countries with different policy portfolios.

This combination of many target and instrument constellations on one hand, and non-incremental policy changes on the other, makes it very difficult if not impossible to assess how much individual policy components contribute to the observed outcome-level changes. Therefore, readjusting or fine-tuning the domestic policy mix becomes increasingly difficult. Not only because policy portfolios are becoming increasingly complex but

also because policy reforms tend to be complex themselves. While the empirical scope of Fig. 4 remains restricted to the fields of environmental and social policy, we believe that these patterns are also reflective of reform trends in other policy sectors.

Discussion and conclusion: trade-offs between methodological sophistication and the political impact of evidence?

While we subscribe to the intention behind evaluating policy mixes in terms of outcome achievements, we used this paper to highlight that outcome measurement is a necessary but unfortunately not sufficient foundation for effective learning about how to improve policy design. The continuous accumulation of policy portfolios makes it more and more difficult to learn about policy effects with conventional approaches.

More specifically, we argued that policy accumulation is becoming an increasingly relevant source of the attribution problem, which consists of assessing the causal link between policy intervention and outcome effect. While conventional discussions of this attribution problem see the source of this problem in the multitude of non-policy factors that influence outcomes, we see policy accumulation as an additional source of this attribution problem that so far has not been systematically addressed. We refer to this accumulation-induced attribution problem as the 'independent variable problem'.

Policy portfolios are not merely complicated systems that comprise a large (and increasing) number of independent policy targets and instruments, but also—and to an increasing extent—reflect complex systems characterized by the interactions between different policy targets and instruments. Complexity is a direct consequence of continuous policy accumulation: as national policy mixes combine a growing number of interacting policy targets and instruments, the degree of policy complexity is increasing. Accordingly, the number of interactions that have to be considered when trying to relate outcomes to existing policy designs is also growing. Accordingly, the number of interactions that must be considered when trying to relate outcomes to existing policy designs is also growing. This complicates our attempts to evaluate the effectiveness of public policies in a way that will be able to feed back into policy-making. To improve the design and composition of ever more complex policy mixes, one must find ways to make statements about the direct and conditional effects of these different parts of the policy mix.

We have shown that the need to estimate conditional effects within increasingly complex policy mixes creates both theoretical and methodological challenges. Bringing all elements within the policy mix to the forefront by explicitly modeling their conditional impacts implies the handling of a high number of parameters, even for relatively simple policy mixes. Depending on the research design and data structure, this can quickly lead to a situation of 'too many variables, too few cases' (Goggin 1986). The resulting challenges to the internal validity of results could be mitigated with the help of quasi-experimental evaluation approaches, which single out individual elements from the policy mix while treating the rest of this mix as constant background factors.

Yet, since the evaluative strategy consists of focusing on policy effectiveness within one specific policy mix, advocates of policy transfer will nevertheless have to deal with the conditionality of the identified effect in the foreign policy mix to anticipate how the policy will perform within their domestic mix. Finally, the political fact that policy reforms are often encompassing rather than incremental further complicates the identification of (conditional) effects of individual policy components within mixes.

Overall, we seem much better equipped to evaluate whether policy mixes as such are associated with improvement of deteriorating outcomes than to assess which elements within these mixes do work effectively and why. This makes it very difficult to reform these mixes in a meaningful way on the basis of evidence, and it leads to a situation that is skewed towards amending existing mixes and against simplifying existing mixes by getting rid of ineffective or inefficient policy elements.

While the empirical focus of our paper is set on environmental and social policy, our argument is not limited to these two sectors. Wherever scholars are confronted with (1) increasing levels of policy complexity, where (2) policy outcomes depend on the interactions between the individual policy components, and (3) policy change is rather encompassing than incremental, they have to deal with this independent variable problem. Of course, this might not be similarly relevant in all policy sectors. Particularly in areas typically considered to be morality policies, such as the regulation of abortion, pornography, and homosexuality, policy accumulation tends to be substantially lower than in areas like labor market policy, or education policy. At the same time, however, outcomes are typically considered much less important in such morality policies.

Some readers might question the relevance of our results, because policy makers rarely form opinions and adopt decisions strictly based on evidence. While it is certainly true that policy decisions are often not based on evidence, the relevance or our contribution lies in the very fact that scholars of public management and public policy criticize this neglect of evidence. Consequently, they promote the stronger institutionalization and guidance of evidence in decision-making processes. We are far from doing the opposite. We share the view that evidence should play a critical role in decision-making. However, we want to use this paper to emphasize that in the light of policy accumulation—a trend that appears to be quite common across developed democracies—creating the kind of evidence we wish to inform decision-making becomes more and more demanding. It is highly questionable whether conventional approaches are able to deal with an aggravating independent variable problem for very long. We thus hope to raise awareness for policy accumulation to be a source of the attribution problem that has so far remained neglected. Our ways of creating evidence must be sensitive towards the analytical challenges that come with continuous policy accumulation, all the more so since reducing policy complexity is hardly a viable option. After all, this complexity is the result of decades and sometimes centuries of efforts to (re)solve problems and conflicts in a legitimate and effective way. Our diagnosis should thus also not be misinterpreted as a call for a return to applying Tinbergen's rule, which states that the optimal ratio of tools to targets in policies should be 1:1. Therefore, this paper should be read as a call for cautiousness and consciousness; as an invitation for a debate about how to institutionalize the kind of evidence-creating systems we desire to inform decision-making.

In this regard, we see a central trade-off between methodological sophistication and the political effects of better and more differentiated evidence. On one hand, there are certainly many ways of further developing and reflecting our methodological toolkit to cope with the challenges of evaluating accumulating policy mixes. While observational research designs appear particularly problematic in the light of policy complexity, they represent an important mechanism that enables us to learn from the experience of others. In addition, although experimental policy roll-outs before policy reforms might not be possible in all areas, this strategy seems to provide an ideal complementary strategy to avoid unwarranted expectations about the effectiveness of policy innovations. It allows for the reassessment of whether policy innovations that proved effective within other policy mixes are able to hold their promises within the domestic or local policy

mix. Only relying on experimental policy roll-outs would probably be insufficient, since such experiments do not really help determine whether innovations to policy mixes introduced and experimentally evaluated elsewhere will yield similar results in other countries with different policy mixes. The systematic combination of observational and experimental research strategies appears more promising in this regard. Furthermore, these triangulation efforts could be complemented by methods from the field of complex adaptive systems and system dynamics. While these approaches have been used to improve performance management and outcome-based decision-making (Bianchi 2016), they neither have become an integral part of mainstream social science curricula, nor the dominant approach used in the political practice of policy evaluation. Likewise, it seems necessary to produce more knowledge about the conditions under which insights gained for a particular policy or policy portfolio are generalizable and can be transferred to other temporal or spatial contexts (O'Toole and Meier 2014). Here, a potential way forward might be to shift the focus from the empirical analysis of the output–outcome nexus to the theoretical examination of the causal mechanisms underlying each policy interventions (Bates and Glennerster 2017). Adopting such perspective, the guiding question must be how the existence (or absence) of other policy measures does affect the "disaggregated theory behind a [policy] program" (ibid.), i.e., the exact reasons and motivations leading citizens to change their behavior as a response to public policies.

While systematic triangulation is—in our view—one potential way forward, the effective implementation of such evidence-generating systems is very demanding. Doing so successfully requires continuous and concerted efforts in data collection as well as methodological expertise in different areas. If we keep in mind that decision makers have been proved to be rather reluctant to implement even less demanding systems, it becomes clear that addressing the independent variable problem requires not only methodological efforts. Even more importantly, it requires substantial political efforts.

On the other hand, even if we assume such political backing, we have to be aware of the consequences of such developments with regard to the political impact of the evidence created. Simply put, when evidence about the effects of specific policy measures is very clear as it portrays a strong and unambiguous message about the obvious advantages and disadvantages of a certain policy, changes in the positions of policy makers should emerge more quickly simply as the trigger is so strong. When 'realistic evaluations' highlight a series of relevant policy factors that condition the effect of individual policies within large and complex policy mixes, the informational trigger is—by definition—not so clear. On the contrary, the informational trigger that is communicated to decision makers is difficult to process. In this situation, posterior beliefs will often differ only slightly from prior beliefs, simply because the informational trigger is blurry rather than crystal clear. Hence, we face a paradoxical situation in which despite the existence of sophisticated and nuanced evidence about policy effectiveness, the impact that this evidence will have on policy makers' thinking is constantly decreasing. Policy accumulation not only drives the need for more—and more sophisticated—analyses and evaluation strategies, it also calls for the identification of highly complex conditional effects. However, the very conditionality and complexity that make these assessments sophisticated stand in the way of having a deep impact on decision makers' thinking.

Acknowledgements This research was funded by CONSENSUS project financed under the European Commission's 7th Framework Programm.

References

Adam, C., Hurka, S., & Knill, C. (2017a). Four styles of regulation and their implications for comparative policy analysis. *Journal of Comparative Policy Analysis: Research and Practice, 19*(4), 327–344.

Adam, C., Knill, C., & Fernandez-i-Marin, X. (2017b). Rule growth and government effectiveness: Why it takes the capacity to learn and coordinate to constrain rule growth. *Policy Sciences, 50*(2), 241–268.

Adam, C., & Raschzok, A. (2017). Cannabis policy and the uptake of treatment for cannabis-related problems. *Drug and Alcohol Review, 36*(2), 171–177.

Albalate, D., Bel, G., & Fageda, X. (2009). Privatization and regulatory reform of toll motorways in Europe. *Governance, 22*(2), 295–318.

Alford, J., & Head, B. W. (2017). Wicked and less wicked problems: a typology and a contingency framework. *Policy and Society, 36*(3), 397–413.

Anderson, D. M., & Rees, D. I. (2013). The legalization of recreational marijuana: How likely is the worst-case scenario? *Journal of Policy Analysis and Management, 33*(1), 221–232.

Babyak, M. (2004). What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine, 66,* 411–421.

Barnow, B. S. (2000). Exploring the relationship between performance management and program impact: A case study of the job training partnership act. *Journal of Policy Analysis and Management, 19*(1), 118–141.

Bassanini, A., & Duval, R. (2009). Unemployment, institutions, and reform complementarities: re-assessing the aggregate evidence for OECD countries. *Oxford Review of Economic Policy, 25*(1), 40–59.

Bates, M. A., & Glennerster, R. (2017). The generalizability puzzle. *Stanford Social Innovation Review, 2017*, 50–54.

Bauer, M. W., & Knill, C. (2012). Understanding policy dismantling: An analytical framework. In M. W. Bauer, A. Jordan, C. Green-Pedersen, & A. Héritier (Eds.), *Dismantling public policies: Preferences, strategies, and effects* (pp. 30–51). Oxford: Oxford University Press.

Bianchi, C. (2016). *Dynamic performance management*. Cham: Springer.

Blonz, J. A., Vajjhala, S. P., & Safirova, V. (2008). *Growing complexities: A cross-sector review of US biofuels policies and their interactions*. Washington: Resources for the Future.

Bovaird, T. (2012). Attributing outcomes to social policy interventions-, gold standard' or 'fool's gold' in public policy and management? *Social Policy and Administration, 48*(1), 1–23.

Brambor, T., Clark, W. R., & Golder, M. (2006). Understanding interaction models: Improving empirical analyses. *Political Analysis, 14*(1), 63–82.

Capano, G., & Lippi, A. (2017). How policy instruments are chosen: Patterns of decision makers' choices. *Policy Sciences, 50*(2), 269–293.

Carley, S. (2009). State renewable energy electricity policies: An empirical evaluation of effectiveness. *Energy Policy, 37*(8), 3071–3081.

Davies, P. (2012). The state of evidence-based policy evaluation and its role in policy formation. *National Institute Economic Review, 219,* 41–52.

Del Río, P. (2014). On evaluating success in complex policy mixes: The case of renewable energy support schemes. *Policy Sciences, 47*(3), 267–287.

Dolowitz David, P., & Marsh, D. (2002). Learning from abroad: The role of policy transfer in contemporary policy-making. *Governance, 13*(1), 5–23.

Eliadis, F. P., Hill, M. M., & Howlett, M. (Eds.). (2005). *Designing government: From instruments to governance*. Montreal, CA: McGill Queens University Press.

Favero, N., Meier, K. J., & O'Toole, L. J. (2016). Goals, trust, participation, and feedback: Linking internal management with performance outcomes. *Journal of Public Administration Research and Theory, 26,* 327–343.

Glennerster, R. (2012). The power of evidence: Improving the effectiveness of government by investing in more rigorous evaluation. *National Institute Economic Review, 219,* 4–14.

Goggin, M. L. (1986). The "too few cases/too many variables" problem in implementation research. *Western Political Quarterly, 39*(2), 328–347.

Gunningham, N., & Grabosky, P. (1998). *Smart regulation: Designing environmental policy*. New York: Oxford University Press.

Gunningham, N., & Sinclair, D. (1999). Regulatory pluralism: Designing policy mixes for environmental protection. *Law and Policy, 21*(1), 49–76.

Head, B. W. (2016). Toward more "evidence-informed" policy making? *Public Administration Review, 76*(3), 472–484.

Heinrich, C. J. (2002). Outcomes-based performance management in the public sector: Implications for government accountability and effectiveness. *Public Administration Review, 62,* 712–725.

Horn, G.-A., & Logeay, C. (2010). Erfolg oder Misserfolg? Die Arbeitsmarktreformen im Rahmen der Agenda. In G. Bäcker, S. Lehndorff, & C. Weinkopf (Eds.), *Den Arbeitsmarkt verstehen, um ihn zu gestalten*. Wiesbaden: Springer VS.

Howlett, M., & del Rio, P. (2015). The parameters of policy portfolios: Verticality and horizontality in design spaces and their consequences for policy mix formulation. *Environment and Planning C: Government and Policy, 33*(5), 1233–1245.

Howlett, M., & Lejano, R. P. (2013). Tales from the crypt: The rise and fall (and re-birth?) of policy design studies. *Administration and Society, 45*(3), 356–380.

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

Howlett, M., & Rayner, J. (2013). Patching vs packaging in policy formulation: Assessing policy portfolio design. *Politics and Governance, 1*(2), 170–182.

Kaplan, R. S., & Norton, D. P. (2004). *Strategy maps: Converting intangible assets into tangible outcomes*. Harvard: Harvard Business Press.

Knill, C., Schulze, K., & Tosun, J. (2012). Regulatory policy outputs and impacts: Exploring a complex relationship. *Regulation and Governance, 6*(4), 427–444.

Linder, S. H., & Peters, B. G. (1989). Instruments of government: Perceptions and contexts. *Journal of Public Policy, 9*(1), 35–58.

Majone, G. (1989). *Evidence, argument, and persuasion in the policy process*. New Haven, London: Yale University Press.

May, P. J. (1991). Reconsidering policy design: Policies and publics. *Journal of Public Policy, 11*(2), 187–206.

Mayne, J. (2007). Challenges and lessons in implementing results-based management. *Evaluation, 13*(1), 87–109.

McBeath, B., & Meezan, W. (2010). Governance in motion: Service provision and child welfare outcomes in a performance-based, managed care contracting environment. *Journal of Public Administration Research and Theory, 20,* 101–123.

McConnell, A. (2010). Policy success, policy failure and grey areas in-between. *Journal of Public Policy, 30*(3), 345–362.

Mettler, S. (2016). The policyscape and the challenges of contemporary politics to policy maintenance. *Perspectives on Politics, 14*(2), 369–390.

Miller, J. H., & Page, S. E. (2009). *Complex adaptive systems: An introduction to computational models of social life: an introduction to computational models of social life*. Princeton: Princeton University Press.

Möhring, K. (2012). The fixed effects approach as alternative to multilevel models for cross-national analyses. *GK SOCLIFE Working Paper Series*, pp 1–15.

Moulton, S., & Sandfort, J. R. (2017). The strategic action field framework for policy implementation research. *Policy Science Journal, 45*(1), 144–169.

Moynihan, D. P. (2005). Goal-based learning and the future of performance management. *Public Administration Review, 65*(2), 203–216.

O'Toole, L. J., & Meier, K. J. (2014). Public management, context, and performance: In quest of a more general theory. *Journal of Public Administration Research and Theory, 25*(1), 237–256.

Pacula, R. L., Powell, D., Heaton, P., & Sevigny, E. L. (2015). Assessing the effects of medical marijuana laws on marijuana use: The devil is in the details. *Journal of Policy Analysis and Management (the Journal of the Association for Public Policy Analysis and Management), 34*(1), 7–31.

Piotrowski, S. J., & Rosenbloom, D. H. (2002). Nonmission-based values in results-oriented public management: The case of freedom of information. *Public Administration Review, 62*(6), 643–657.

Pirog, M. A., & Ziol-Guest, K. M. (2006). Child support enforcement: Programs and policies, impacts and questions. *Journal of Policy Analysis and Management, 25*(4), 943–990.

Pollitt, C. (2011). Performance blight and the tyranny of light? Performance blight and the tyranny of light? Accountability in advanced performance measurement regimes. In M. J. Dubnick & H. G. Frederickson (Eds.), *Accountable governance: Problems and promises* (pp. 81–98). M.E. Sharpe: Armonk.

Radin, B. A. (2000). The government performance and results act and the tradition of federal management reform: Square pegs in round holes? *Journal of Public Administration Research and Theory, 10,* 111–135.

Radin, B. A. (2006). *Challenging the performance movement: Accountability, complexity, and democractic values*. Washington, D.C.: Georgetown University Press.

Radin, B. A. (2009). What can we expect from performance measurement activities? *Journal of Policy Analysis and Management, 28,* 505–512.

Rinne, U., & Zimmermann, K. (2012). Another economic miracle? The German labor market and the Great Recession. *IZA Journal of Labor Policy, 1*(1), 1–21.

Rittel, H. W. J., & Webber, M. M. (1973). Dilemmas in a general theory of planning. *Policy Sciences, 4,* 155–169.

Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2003). *Evaluation: A systematic approach*. Thousand Oaks: SAGE Publications.

Schickler, E. (2001). *Disjointed pluralism: Institutional innovation and the development of the US congress*. Princeton: Prrinceton University Press.

Schlaufer, C., Stucki, I., & Sager, F. (2018). The political use of evidence and its contribution to democratic discourse. *Public Administration Review*. https://doi.org/10.1111/puar.12923.

Schneider, A. (2012). Policy design and transfer. In E. Araral, S. Fritzen, M. Howlett, M. Ramesh, & X. Wu (Eds.), *Routledge handbook of public policy* (pp. 217–228). Abingdon: Routledge.

Thelen, K. (1999). Historical institutionalism in comparative politics'. *Annual Review of Political Science, 2,* 369–414.

Thelen, K. (2004). *How institutions evolve: The political economy of skills in Germany, Britain, the United States, and Japan*. Cambridge: Cambridge University Press.

Van Dooren, W., & Van de Walle, S. (2016). *Performance information in the public sector: How it is used*. Berlin: Springer.

Van Thiel, S., & Leeuw, F. L. (2002). The performance paradox in the public sector. *Public Performance and Management Review, 25,* 267–281.