



Causality is good for practice: policy design and reverse engineering

Simone Busetti¹

Accepted: 20 January 2023 / Published online: 1 February 2023
© The Author(s) 2023

Abstract

Relevance to practice is an open issue for scholars in public policy and public administration. One major problem is the need to produce knowledge that can guide practitioners designing and implementing public interventions *in specific contexts*. This article claims that investigating the causal mechanisms of policy programs—i.e., modeling why and how they produce outcomes—can contribute to such knowledge. In this regard, mechanisms offer essential information to guide practitioners when replicating, adjusting, and designing interventions. Unfortunately, not all models of mechanisms can inform practice. The article proposes a strategy for design research and practice inspired by reverse engineering: selecting successful programs, causal modeling, assessing the target context, and designing. Scholars should model mechanisms by identifying the program and non-program elements that contribute to the outcome of interest and abstracting their causal powers. Practitioners can use these models, diagnose their target context, and adjust designs to deal with context-specific problems. The proposed research agenda may enhance orientation to practice and offer a middle ground between the search for abstract, general relationships, and single-case analyses.

Keywords Causal mechanisms · Policy design · Policy implementation · Usable knowledge · Practitioner

Back to practice

Relevance to practice is at the root of public policy and public administration. In a foundational article, Harold Lasswell described the distinctive outlook of the policy scientist as one comprising contextuality, problem orientation, and a distinctive synthesis technique (Lasswell, 1970). Based on the unity of theory and action, policy inquiry was aimed explicitly at solving problems in specific times and places (Lasswell & Kaplan, 1950). In a similar vein, public administration was viewed as practice-oriented, advancing the knowledge

✉ Simone Busetti
sbusetti@unite.it

¹ Department of Political Science, Università degli Studi di Teramo, Via Renato Balzarini 1, 64100 Teramo, Italy

of public affairs and the work of public administrators: it entailed both study and action (Waldo, 1955).

Calls for usable knowledge (Lindblom & Cohen, 1979), a social science that matters (Flyvbjerg, 2001), and a paradigm for practice (Brunner, 2006) indicate the ongoing attention paid to enhancing practical relevance. However, they also point to an enduring dissatisfaction with the fact that social science is at risk of becoming “a sterile academic activity, in increasing isolation from a society on which it has little effect” (Flyvbjerg, 2001: 165).

Indeed, several factors counter practice-oriented scholarship. A partial list would include a rationalistic bias ignoring political problem-solving (Lindblom & Cohen, 1979), reductionism and the search for general relationships (Brunner, 2006), context-independency and the revered role of prediction (Flyvbjerg, 2001), an axiology of science hindering applied knowledge (Pielke, 2004), a systematic bias against interdisciplinary research (Raadschelders, 2011), and concerns about the status and careers of practice-oriented researchers (Bailey, 2006).

One commonly encountered point concerns the difficulty in producing practice-oriented knowledge, i.e., knowledge that can guide practitioners when they design and implement public interventions *in specific contexts*. Professionals deal with particular situations—individual cases, classrooms, or communities—and need adequate knowledge (Stake, 2010). The article discusses how investigating the causal mechanisms of the program—i.e., modeling *why* and *how* program and non-program elements produce the outcome of interest—is a fundamental piece of knowledge contributing to this endeavor.

Under the label of causal mechanisms, the social sciences have witnessed much interest in the analysis of causality not only in policy analysis, public administration, management, and evaluation (Bardach, 2004; Barzelay, 2007; Buseti & Dente, 2018; Capano & Howlett, 2019; Melloni et al., 2016; Ongaro, 2009; Pawson & Tilley, 1997) but also and primarily in sociology, political science, and methodology (Beach & Pedersen, 2018; Bennett, 2013; Falletti & Lynch, 2009; Gerring, 2010; Mahoney, 2001; Mayntz, 2004; McAdam et al., 2004).

Although there is almost unanimous agreement that causal mechanisms improve explanations and reduce the risk of spurious relationships (see Bunge, 2004; Mahoney, 2001), the literature is less compelling regarding whether such “superior” explanations have practical utility. In fact, there have been claims that this knowledge may well be unnecessary, specifically in the case of policy programs, for which “the covariational relationship is usually the key component of concern” (Gerring, 2010: 1506).

The first goal of this article is to challenge this view. Causal mechanisms have a great deal of practical utility; they provide essential information for use in designing more effective interventions. The first section presents three types of knowledge that together inform practice: evidence about effectiveness, information about program implementation, and causal mechanisms. The second section proceeds by reviewing how knowledge of causal mechanisms can support a wide range of design activities: intelligent replications, program adjustments, and the design of brand-new interventions. If scholars in public administration and policy analysis want to increase their relevance to practice, they need to engage in theory-building research and improve our understanding of how and why policies work; to this end, mechanisms should take center stage in their research agendas.

The second goal of this article is to provide suggestions regarding how to investigate mechanisms with a practice orientation. In fact, not all references to causal mechanisms will deliver usable information: they may improve explanations without helping practitioners with their context-specific design and implementation problems. The fourth section stresses that causal mechanisms are not obvious; they can neither be inferred deductively

nor assumed *ex post* but should, rather, be researched purposefully. Then, the fifth section presents a strategy of design research and practice inspired by reverse engineering. This strategy includes four activities, each discussed in one of these sections: Selecting successful programs, Causal modeling, Assessing the target context, and Designing programs in new contexts. The first two activities concern research. In particular, the “[Causal modeling](#)” section allows the specification of a usable definition of mechanisms; in order to inform practice, models of causal mechanisms need to track the process by which program and non-program elements contribute to outcomes and investigate the causal power of those elements. The sections “[Assessing the target context](#)” and “[Designing programs in new contexts](#)” concern practice; they respond to the third objective of this article: providing suggestions on how to use mechanisms for improving policy design and implementation in specific contexts. Finally, the conclusions discuss the wider utility of this approach.

Three types of practical knowledge

Imagine a public official designing a new program. Besides analyses of the policy problem, there are at least three kinds of knowledge that can help in crafting and implementing an effective intervention.

The most obvious relates to evidence about existing programs that are effective in solving the same problem. For instance, if in a school district, students are experiencing poor achievement, the designer will look for a program with a record of effectiveness in improving student results, perhaps finding experimental evidence that class size reduction (CSR) improves skills (Finn & Achilles, 1990) or statistical evidence that identifies a negative and significant relationship between the number of students per teacher and the pass rate for a certain standardized test (Gill & Meier, 2001). Following evidence-based policy, this information can be qualified with data about when, where, and for whom the program is working (Davies et al., 2000) and serve as a fundamental starting point; now, the practitioner knows of a program somewhere that worked and can draw the following clear, practical advice: implement a similar program. This kind of information will ultimately offer practitioners a range of design options or a menu of policy tools and instrument mixes for solving the problem.

However, testing effectiveness says little about how to implement a program in the practitioner’s context. A second type of knowledge that the practitioner will certainly need is descriptive information about how to design and implement the selected intervention. Reducing the number of students entails a huge number of decisions that are anything but trivial regarding the funding needed to deliver the program, the hiring policy for new teachers, facilities, as well as information about any legal, administrative, or organizational issues involved in setting up the program. Academic research typically disregards these details to the point of focusing on a single feature, namely the reduction of the number of students, which is insufficient for putting the program into practice. This second type of knowledge is indispensable and typically circulates in formal and informal networks of practitioners.

Finally, knowledge of causal mechanisms can offer an additional piece of information for practice. Evidence that a program is effective (e.g., CSR is a cause of student improvement) does not tell us why and how that program produces those results. In the case of CSR, there exist at least two hypotheses regarding why students’ scores may improve: because teachers change their teaching strategies or because smaller classes neutralize

social loafing, i.e., students' propensity to hide in large groups (Finn & Achilles, 1999; Finn et al., 2003). The next section discusses how this knowledge is essential for designing and implementing programs.

Design advantages of causal mechanisms

The first advantage of investigating mechanisms is that they improve our understanding of what it is about an intervention that contributes to producing results. Imagine that after digitizing a certain administrative procedure, there is a drop in the number of days needed to complete that procedure. Let us say that new information and communication technology (ICT) allows the automatic execution of some activities and facilitates the transmission of documents between the offices involved. The ICT mechanism may be as simple as this: streamlining procedures and enhancing organizational efficiency. Let us say, however, that the same features that allow the transmission of documents also increase transparency—i.e., everyone knows which office is processing a given case and how long it is taking—and that greater speed is achieved because employees give priority to that procedure to avoid being considered laggards. It is only when the researcher discloses this latter mechanism that the designer knows what it is about the ICT that makes it work. This information can bring about a complete change of perspective and reorient the designer's job more precisely toward what the ICT is supposed to accomplish (streamlining vs. prioritizing).

The same reasoning applies to the relevant “non-program” elements, i.e., those that are not designed as part of the program but contribute to making it work (for instance, implementation capacity). Consider the case of CSR again. If the program works by neutralizing social loafing, the non-program element of interest is the reaction of the students; if, instead, CSR produces a change in teaching techniques, it is teachers who are contributing to change. Again, these differences alter the designers' focus and design activity completely. The same occurs if the designer knows that greater speed is not the automatic product of digitization but rather requires a behavioral change on the part of implementers (e.g., assigning priority)—a change that cannot be taken for granted in all contexts.

This knowledge does not simply improve our explanations; it is essential for practice. Three such design advantages are presented in the next subsections: intelligent replications, program adjustments, and the design of new programs. Notice that while the importance of mechanisms for replicating programs has been discussed elsewhere (Bardach, 2004; Barzelay, 2007; Cartwright & Hardie, 2012; Williams, 2020), their wider relevance to policy design is still little appreciated.

Intelligent replications

A major problem for practitioners concerns how evidence about effectiveness in one place (our first type of knowledge) can be used in the implementation of a similar intervention in another context. Unfortunately, “the cleanest estimation of a program's impact does not provide warrant for confidently inferring that similar results can be expected if that project is scaled up or replicated elsewhere” (Woolcock, 2022). Interestingly, instead, knowledge about causal mechanisms is said to prevent the risk of naïve transfers and errors in “learning from second-hand experiences” (Barzelay, 2007): mindlessly reproducing surface

features, neglecting relevant details, and disregarding contextual differences (Bardach, 2004; Barzelay, 2007; Buseti & Dente, 2021; Cartwright & Hardie, 2012; Ongaro, 2009).

The first two problems are straightforward. Even simple programs have a wide variety of features; some are fundamental to solving the policy problem and must be replicated in the new context, while others may be irrelevant or even dysfunctional and may be disregarded. Without knowledge about mechanisms, however, it is difficult to know which is which. In the ICT example, a designer who does not know that the procedure works by enhancing transparency may well reproduce the program without those apparently minor details that allow everyone to check where the document is being processed.¹

The third problem in replicating programs concerns contextual fitness. In a different context, an effective intervention may not find the same conditions that originally contributed to its success. Cartwright and Hardie (2012) provide a telling example in the form of the transfer of a nutrition project from Tamil Nadu to Bangladesh. One of the measures included in the original program was nutritional education for pregnant mothers. When exported to Bangladesh, however, this measure was ineffective because mothers were not in control of food in the new context: men, not women, go to the market in rural Bangladesh, and when mothers-in-law live with the family, they are the ones who govern the house.

The obvious lesson is that due to contextual discrepancies, the mere transfer of the program may fatally fail.² The less obvious lesson is that if one moves past the simple description of the program (i.e., educating mothers) and extrapolates why it works (i.e., by changing the behavior of the person who manages the children's diet), problems of contextual fitness may be predicted and avoided (for instance, by including mothers-in-law in the program). Intelligent replications do not entail making exact copies of the original program but rather redesigning it to conserve the original mechanism and possibly obtain the same results.

Program adjustments

Heterogeneous results do not happen only when replicating programs; they are also the norm in the original context where success was first detected. The average results of the program will mask a varied reality of excellent, good, sufficient, and poor performers. Indeed, while researchers with theoretical concerns may be satisfied with performance that is effective on average (Meier & Gill, 2000), practitioners require knowledge about what to

¹ As suggested by one reviewer, if there are different hypotheses about what is causally active in the program, multiple RCTs could test these hypotheses. Although in principle this is certainly possible, there are several complications in pursuing this strategy. Woolcock (2022) suggests that RCTs are possible only when there is a limited number of elements interacting in predictable ways, the effect of particular elements can be isolated, and the “black box” has basically been opened. In the same vein, Green, Ha, and Bullock (2010) claim that RCTs would need too strong requirements in model specification in order to test mediators (e.g., whether the same ICT works by streamlining or transparency). Finally, and more fundamentally, without having a “transparency hypothesis” in the first place, there is no way to test it.

² These problems are not specific to any method. They are discussed in the case of RCTs (Cartwright & Hardie, 2012), as well as in field experiments and pilot interventions, where the original success may incur a “voltage drop” (Al-Ubaydli et al., 2019). Brunner (2006), Schön (1983), and Flyvbjerg (2001) note the insufficiency of inferences from statistical models when applied in specific contexts. Overman and Boyd (1994) raise a similar point regarding case studies that describe “a-theoretical best practices”—i.e., descriptive accounts of successful programs that, although originally effective, are hardly transferable across contexts.

do in specific cases, such as when, in their school district, a program that is successful on average at the national level is not working or is underperforming. In this latter case, one may develop several options that are known to be positively correlated with student results: implementing a larger reduction in the number of students per class, adding another grade of small classes, improving facilities, and introducing special arrangements for problematic students. On average, all these options may certainly improve the performance of problematic districts. As such, however, they will be selected without a well-grounded idea about how to precisely strengthen why small classes work in the first place. A more focused attempt may entail beginning with the causal mechanisms of CSR and then devising changes to the original design that support and reinforce those mechanisms in that context. According to the hypothesis of innovative teaching strategies, for instance, this may occur via training teachers and actively promoting a change in teaching techniques, in addition to simply reducing the number of students. If social loafing were taking place, different adjustments would be in order.

If success is not homogeneous, is it also not static. Implementing a program is a permanent activity. Things may change in the implementing administration, in the target group and beneficiaries, or in a huge number of contextual conditions. When things change, even a successful program may lose its grip, and results will deteriorate. Once again, the practitioner will require knowledge about mechanisms to adjust the program in a focused way.

Suppose, for instance, that the initial results obtained with the ICT procedure begin declining—let us say because, after the initial period, employees become used to this procedure and stop over-performing. A designer who knows that the ICT makes employees prioritize that procedure may decide to stress that behavior purposefully, for instance, by introducing an explicit tracking device or including some nudges in the exposure of bad performance. These adjustments will not be generic improvements to the program, such as adopting a high-performing technology, but will instead build on the original mechanism triggered by the ICT.

More generally, knowledge about causal mechanisms may support decisions on all sorts of program adjustments, even those regarding scaling the program. Given the success of the ICT procedure, it would seem natural, or even imperative, to extend it to all procedures in the administration. But would that be a good idea? Without mechanisms, the answer is indeterminate. If the ICT solution works by streamlining procedures, an extension would certainly speed up other procedures as well; however, if it causes employees to give priority to the tracked applications, a universal extension will neutralize its effect altogether (by definition, there is no such thing as a universal priority).

The examples could go on, but the general point is that—viewed against the huge variety of choices that the practitioner will be confronted with in practice—mechanisms provide a compass with which to evaluate such choices and design adequate responses.

Designing new programs

Ensuring success may require more than simple strengthening or fine-tuning. In some cases, even the core components of the selected intervention will not be viable. Let us imagine the most dramatic situation in the CSR example: in some districts or schools, it is impossible to reduce the number of students to a significant degree (let us say due to a lack of facilities). What can the practitioner do? Even in these situations, causal mechanisms may suggest intelligent responses.

The designer may begin by considering why CSR works and devising an alternative that conserves the same mechanism—such as a new program or design feature that neutralizes social loafing notwithstanding large classes. For example, the designer might try the “flipped classroom”: recording lectures as homework assignments and using all face-to-face classroom time for interactive learning (Missildine et al., 2013).

More generally, the practitioner will redesign the selected program by introducing “functional equivalents” (Rose, 1993), i.e., alternative or additional design features triggering the same causal mechanism. Consider a program for improving the diet of students by distributing apples at school. In case apples are unavailable, could other fruits be functional equivalents? This might seem like an easy question because one implicitly assumes that whatever the mechanism triggered by the apple, it is something that other fruits can do. Imagine, however, that the apple has some specific causal feature, such as the ability to remain intact in a school bag, travel safely home, and serve as a parental memo for recommending one fruit a day. In this case, not all fruits or delivery methods are functional equivalents, and it is only by disclosing the mechanisms of apple distribution that the designer can discern which are. The same reasoning applies *a fortiori* to programs more complex than fruit distribution and design elements more sophisticated than apples.

Importantly, the utility of mechanisms for design is not limited to the same problem or sector. Once the researcher has extrapolated the mechanisms of a program, that knowledge is part of the designers’ toolkit and will be used whenever needed for crafting new interventions. In the ICT example, the increase in transparency produces speed because employees give priority to the tracked application to avoid the blame for underperformance compared with their peers. Melloni (2013) identifies the very same mechanism to explain a different outcome, i.e., the high quality of EU impact assessments elaborated by the EU Directorates-General. In her case, the procedures for drafting and discussing the assessments differ from the ICT example but still promote transparency and peer review and trigger a “blame-avoidance” reaction by EU bureaucrats (Melloni, 2013). Buseti and Dente (2018) model the mechanism of “blame avoidance” by referring to the interplay of three elements: design features enhancing transparency, peers attributing blame, and blame-sensitive subjects changing their behaviors. Using this abstract model, one may design ICT tracking to speed up administrative permits, inter-directorate meetings for improving the quality of EU impact assessments, or any other system of horizontal monitoring that is well suited to the practitioner’s context.

Program causality is not self-evident

If mechanisms are in fact practical, another major misconception is that they can be easily inferred deductively *ex ante* (by simply observing the design of the policy) or assumed *ex post* after having appraised the effectiveness of the program. In fact, even standardized interventions may entail complex and unclear mechanisms that must be uncovered through specific research.

Pawson and Tilley (1997) present an apt example concerning the use of CCTV cameras to protect against car theft. They hypothesize eight mechanisms involving phenomena as varied as the deterrence of thieves, the enhanced operational or investigative skills of the police, and the reaction of car owners. If this sounds remarkable for a simple piece of technology such as a CCTV camera, researchers are advised not to underestimate the causality of any intervention. Notice, however, that the uncertainty arising from having multiple

hypotheses and discriminating between plausible mechanisms (a problem of theory-testing) (Hedstrom and Ylikoski, 2010: 54; Steel, 2004: 65) is only part of the issue; researchers should especially appreciate the possibility of having only wrong hypotheses—hence the importance of engaging in dedicated research into model mechanisms (a problem of theory-building). Two points highlight the urgency of this research.

First, policy programs are not crafted to be purely “functional.” The textbook design recipe imagines designers who craft artifacts with the sole aim of performing a given function (Gero, 1990; Tjalve, 1979). This recipe assumes that there is a clear *ex ante* hypothesis concerning why and how each detail included in the final design will contribute to the expected results. In the case of policy programs, however, “non-functional” processes, such as log-rolling, conflict, and bargaining, have a great deal to say about how policies are ultimately designed (Bobrow, 2006). Programs are path-dependent (Kay, 2005), interventions and institutions are layered one over the other (Ackrill & Kay, 2006; Capano, 2019), and several contextual conditions constrain the scope of the design process (DeLeon, 1988; Dryzek & Ripley, 1988). Briefly put, non-functional factors are standard ingredients in policy formulation, a process that cannot be equated with matching solutions to problems. Contrary to the textbook recipe, why and how the resulting design will work will hardly be clear *ex ante*.

Second, policy programs do not work in isolation. On the contrary, non-program elements are “constitutive” of a program (Barzelay, 2019: 109). Long ago, the implementation literature pointed to the insufficiency of designs in explaining results, and systematic reviews of relevant factors amassed hundreds of additional success variables (O’Toole, 1986, 2000). Any satisfactory account of the mechanisms of a program should certainly consider a configuration of factors beyond design elements—a configuration that can hardly be apparent *ex ante*.

Reverse engineering: research and practice

For the reasons explained in the previous section, mechanisms should never be taken for granted but rather explicitly investigated. Researchers should engage in theory-building research to model why and how program and non-program elements contribute to producing the outcome; then, designers will have a range of models available and use them to solve context-specific problems. This strategy recalls “reverse engineering” (Barzelay, 2019; Busetti & Dente, 2018; Chikofsky & Cross, 1990; Weaver, 2019), a design technique that inverts the textbook design process. Instead of beginning with an existing social problem and designing a “solution,” reversing entails investigating existing policy programs, modeling their mechanisms, and then using these models as a guide for the design process. This strategy is presented in the next four subsections: Selecting successful programs, Causal modeling, Assessing the target context, and Designing programs in new contexts (see Fig. 1).

Notice that the reference to reverse engineering does not imply that the approach is limited to replicating existing programs, nor that the four activities are to be completed together in one sequence. In fact, they refer to two different moments: one regarding research and one regarding practice. Selecting successful programs and causal modeling concern the production of design knowledge in the form of models of causal mechanisms. These models contribute to the advancement of policy design; they constitute a range of options to be included in the designer’s toolkit. The second two activities—assessing

Fig. 1 Reverse engineering: research and practice

RESEARCH	SELECTING THE PROGRAM <ul style="list-style-type: none"> Investigate successful programs in order to reproduce their outcomes
	CAUSAL MODELING <ul style="list-style-type: none"> Use a varied range of evidence Abstract causal powers Include program and non-program elements Draw a model
PRACTICE	ASSESSING THE CONTEXT <ul style="list-style-type: none"> Use the model to check for discrepancies and problematic contextual conditions
	DESIGNING <ul style="list-style-type: none"> If needed, introduce functional equivalents or add design features Change design to fit context and conserve the original mechanism

the target context and designing programs in new contexts—are the work of practitioners. When faced with a policy problem, designers will draw on available models of causal mechanisms to fulfill their design tasks: replicating, adjusting, or creating a brand-new program. In this effort, they are called upon to use their skills, knowledge, and experience in order to fit the modeled mechanisms, design features, and target context.

Selecting successful programs

Researchers should reverse-engineer positive deviants, i.e., outliers that exhibit superior performance relative to a reference group (Bradley et al., 2009; Cammett, 2022). In the case of policy programs, success is a multifaceted concept that can be defined according to several dimensions: programmatic (i.e., producing achievements), process (i.e., the appropriateness of policy-making and implementation), political (i.e., raising social, political, and administrative support), and endurance (i.e., sustained results over time) (McConnell, 2010; Luetjens, Mintrom, and’t Hart 2019; Lindquist et al., 2022). The selected cases do not need to satisfy all possible dimensions of success nor every indicator in one of those dimensions. The researcher will start with the specific outcomes of interest and select high-performing cases accordingly.

The intuitive reason for using positive deviants is that designers should pick something worth replicating. Positive deviance assumes that knowledge about what works already exists in such cases that are delivering good results, while it is absent in the case of failures (Bradley et al., 2009). In fact, successful interventions convey most policy-relevant information (Meier & Gill, 2000) and allow the researcher to trace the conditions of success (May, 1992). Deviant cases are especially suitable for the purpose of modeling the mechanisms of successful cases; they may provide new explanations, update existing models (Gerring, 2007), and support a broad range of discovery-related goals, such as collecting

new information about causal pathways, identifying omitted variables, finding unknown causal paths, and uncovering neglected sources of causal heterogeneity (Seawright, 2016). These advantages are particularly relevant when—starting with uniform designs across contexts—one needs to reveal those features that explain the superior performance of the selected cases (Bradley et al., 2009).

Causal modeling

In the analysis of the selected program, the researcher will work iteratively between formulating hypotheses and collecting data to progressively refine a causal model of the program's mechanisms. Causal modeling broadly follows the method of theory-based evaluation (Weiss, 1997) and is inspired by realistic approaches, especially regarding the attention paid to the interplay between design and context (Pawson & Tilley, 1997). It also borrows from program theory (Funnell & Rogers, 2011) and process tracing (Beach & Pedersen, 2016), in particular with respect to the importance of empirically tracing the hypothesized mechanism and explicitly drawing a causal diagram representing the modeled mechanism.

In light of the variety of methods, techniques, and pieces of evidence that may converge in modeling the program's mechanism,³ three activities are in order to ensure that these models are informative and usable: abstracting causal powers, including multiple elements, and assembling the model.

Abstracting causal powers

The first step is to investigate *why* the program contributes to the outcome. To this end, a useful concept is that of generative causation, developed within the realm of scientific realism (Bhaskar, 2013). The idea is that objects have inherent causal properties or inner causal powers; they can produce certain effects by virtue of characteristics that enable the production of those effects. Gunpowder explodes because of its chemical instability; people can walk but cannot fly because of their anatomy, musculature, density, and shape (Sayer, 1992); the ICT procedure increases speed because its design features have the power to enhance transparency.

Starting with concrete phenomena and their descriptive features, one should abstract their causal power by answering “*why*-questions” such as the following: What is it about the object that makes it do such and such (Sayer, 1992)? By what means does it work? In virtue of what does it produce these results (Cartwright & Hardie, 2012)? In the case of policy interventions, it may be helpful to start with the subjects potentially affected by the program and ask what it is about the intervention that may influence their behavior and

³ As put by Bunge (2004), there is no specific method for conjecturing mechanisms. Case studies are said to have a recognized advantage for theory generation (Gerring 2007), and some guidelines exist regarding conducting theory-building case studies using process tracing (Beach and Pedersen 2016). However, virtually all methods may provide useful insights. In the case of CSR, for instance, evidence may be derived from participant observations of classes (Graue et al., 2007), surveys (Betts and Shkolnik 1999), experiments (Finn and Achilles 1999), regression models (Levin 2002), and path analysis (Bourke 1986). Meier and Gill (2000), for instance, suggest a mixed method; they use weighted regressions to identify the factors that distinguish the best school districts (e.g., teacher certifications, state aid, and CSR) and then propose in-depth case studies to inquire into exactly what these districts are doing with those factors (Meier and Gill 2000).

produce the outcomes of interest. In the CCTV example, for instance, there are three potentially relevant subjects: car owners, police, and thieves. What is it about CCTV cameras that may affect these actors in a way that contributes to decreasing thefts? Starting with the police, for instance, the researcher may formulate hypotheses and speculate regarding whether the causal power of cameras lies in collecting data on criminal practice or allowing timely interventions.

In this process, the practical importance of abstracting causal powers from the simple description of the program cannot be emphasized enough. “Educated mothers” may certainly be the cause of the improvement in children’s diet and fully explain why the Tamil Nadu program works. Abstracting from “mother” (a description) to “person with the power to change the children’s diet” is vital, however. It is the necessary step in extracting causal knowledge that can travel across contexts and be used when transferring, adjusting, or inventing programs.

In order to be usable, however, such abstraction must be neither too general nor tautological. If the researcher says that an intervention works because it provides an incentive, he is abstracting its causal power but in a way that is so general as to provide little practical information (virtually all programs provide incentives—a sanction, a piece of information, money, or other resources—that encourage behavioral change). Usable causal powers must be specific. All campaigns about energy consumption, for instance, give some kind of information, but to generate practical insights, the researcher should abstract what it is about the campaign that makes it work. This may be because it precisely identifies the behaviors to be avoided or because it suggests easily implementable tips. These latter options exemplify what Cartwright (2007) calls “thick causal concepts,” i.e., the variety of concrete causal relationships that provide useful information that is dramatically suppressed by using abstract verbs such as “causing” and “preventing.”

A similar point is made by Sayer (1992) when he warns against causal tautologies, such as the incentives that have the power to incentivize or information campaigns that have the power to inform; tautologies that convey no policy-relevant information. The suggestion is to empirically establish what it is about the object of study that gives it its specific causal power, a causal power to be identified independently of the mere exercise of that power (Sayer, 1992: 72). In the case of the information campaign to reduce energy consumption, “giving tips that are easily implementable” is a practical (and thick) causal power: it does not merely describe the campaign, and it is neither too general nor tautological.

Including multiple elements

A model of a mechanism will necessarily include elements beyond design features. The fact that multiple elements contribute to explaining social phenomena is well-rooted not only in policy analysis but in the overall methodological debate about causality in the social sciences. Causality lies in “sets of interacting components” (Steel, 2008), “cogs and wheels” (Elster, 1989), “causal recipes” (Ragin, 2008), and “causal cakes” (Cartwright & Hardie, 2012); policy interventions only “contribute” to results, together with a multitude of other factors (Mayne, 2012). These suggestions are congruent with a widely accepted definition of causal mechanisms, according to which a mechanism specifies the intervening causal process or pathway, i.e., *how* certain initial conditions produce the outcome of interest (see Hedström & Ylikoski, 2010; Gerring, 2008; Mayntz, 2004).

Following scientific realism, causal powers are not exercised deterministically but rather can remain “dormant” until they are activated in the presence of other elements also

provided with inherent causal powers (Collier, 1994; Sayer, 1992). The power to change behaviors in the ICT example is inherent in the way the procedure is designed, i.e., its capacity to enhance transparency and allow peer review. Still, this power will only be exercised by implementers provided with sensitivity to the public disclosure of their behaviors. The same applies to CSR, which always reduces the number of students but will not have an impact if teachers or students are resistant to change. No program actually works *ceteris paribus*.

The *why*-questions used to understand the causal power of designs also help generate hypotheses for which non-program elements contribute to the outcome. If CCTV cameras help collect data on criminal practice, for instance, the researcher may inquire into what else is needed to support this causal power, such as the capacity to collect and process these data and elaborate targeted strategies to combat thefts. A related suggestion is to trace the intervening process from design to outcome. This can be worked forward (from the causal power of design) or backwards (from the outcome) to help suggest hypotheses on what is supposed to happen—and which elements are involved—in producing the outcome.

In general, these will be those typically identified by the policy literature, such as design and implementation features, characteristics of the target population, and contextual conditions (Sabatier & Mazmanian, 1980; Van Meter & Van Horn, 1975). However, whatever factors the literature and the empirical analysis suggest for inclusion, it is always essential to explicitly research their causal powers. Two examples may serve to illustrate this point. First, implementers and beneficiaries are not passive actors (Pawson, 2006) and, unless they are provided with some relevant causal power (e.g., sensitivity to blame in the ICT example), may not react (or not as expected) to design features. This is self-evident when target groups have interests contrary to the goals of the policy (e.g., the thieves in the case of the CCTV), but it is worth remembering even when their preferences or duties are assumed to be consistent with it. Second, contextual variables are often considered without explicit reference to their causal power, but this limits the utility of contextual information for policy design. For example, time availability is known to be widely relevant in explaining policy success (Durant, 1984; Hogwood & Gunn, 1984). Leach and Pelkey (2001) published a systematic review of the factors explaining the success of watershed management partnerships, and “having sufficient time” stands out as the most relevant variable in the analysis. This may appear self-evident, but why is time relevant? Is it because watershed management requires long-run monitoring and gradual fine-tuning? Does time help participants develop mutual trust? Without this piece of information, practitioners will not know how to use the knowledge that time is relevant, especially when time does not abound in their partnership and they are looking for ways to make it work anyway.

Assembling the model

The final result will be a model of the hypothesized mechanisms representing the causal path from design to outcomes, the different elements contributing to results, and their causal powers. The model can take the form of a causal diagram, i.e., a graphic model portraying the assumptions about the hypothesized causal relations (Greenland et al., 1999; Pearl & Mackenzie, 2018). Diagrams allow us to be explicit about assumptions; their utility for measuring causal effects has been widely acknowledged (Greenland et al., 1999; Pearl, 1995; Pearl & Mackenzie, 2018), and they are also a good practice for informing designers. Figure 2 shows the progress from a black-boxed relationship (A) to the mechanism of the ICT (B) and a more abstract model of horizontal monitoring (C). Although it is

a simplified and hypothetical example, it nonetheless helps summarize the points discussed thus far.

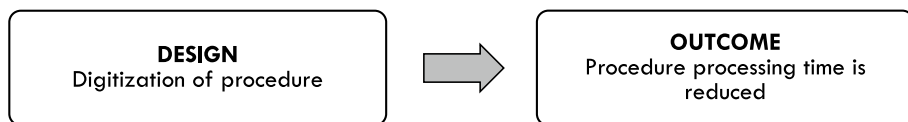
First, the models in B) and C) represent both program and non-program elements (e.g., peers' reactions) and state their causal powers explicitly, i.e., *why* they contribute to the outcome. Peers are provided with some propensity to blame (so as to control and reprimand poor performance) and implementers with some sensitivity to shame (in order to react by increasing their commitment). Precisely because causal powers are explicit, the arrows represent the time sequence but do not obscure the causal linkages between the elements in the model. In this regard, the two diagrams are more informative than one that simply groups together a configuration of relevant variables or one derived from logic models and descriptive theories of change representing the stages of the intervention and the involved inputs, activities, outputs, and intermediate outcomes. If they do not inquire into causal powers, these graphic representations miss fundamental information for designers.

Second, the diagram should illustrate *how* design and non-design elements are supposed to produce the outcome and be explicit about the many assumptions supporting this path. The researcher must open the black box between designs and outcomes and ensure “productive continuity,” i.e., make connections intelligible and avoid gaps in the causal path (Machamer et al., 2000). This explicit identification of the causal process has practical value. In the case of the ICT, for instance, referring only to “blame avoidance” would certainly hint at an explanation, but as such it would stress only one element of the mechanism (the reaction of the sanctioned subject who avoids blame) and black-box others that are equally fundamental for policy design. These practical limitations exist with unspecified references to causal mechanisms that are often present in the literature in the form of synthetic causal labels, such as “power” (Rueschemeyer, 2009), “coordination” and “rational choice” (Falleti & Lynch, 2009), or “culture” and “negotiation” (van der Heijden et al., 2019).⁴

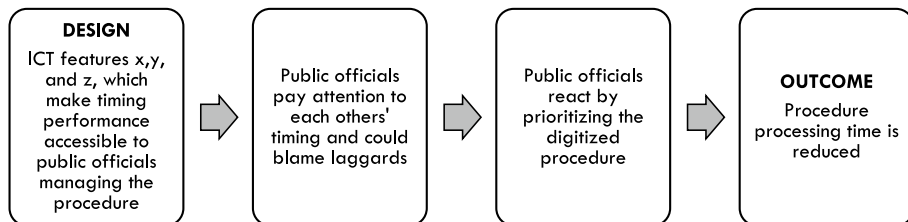
Finally, mechanisms can be modeled at different levels of abstraction. The diagram in B) explains how the ICT works and can be used for new applications or for adjusting and maintaining the program in order to ensure the reduction of procedural time. It incorporates essential information about the process and about which program features trigger the mechanism. The model in C) is more flexible; the designer can use it across contexts, sectors, and problems for designing new programs that enhance performance through horizontal monitoring. This abstract model does not specify any particular design feature but informs—no matter which design is implemented in the target context—that it should be arranged in such a way as to give it the causal power of increasing transparency. Similarly, the abstraction of the implementers' behavior from “prioritizing the digitized procedure” to “increasing commitment” makes the model adjustable toward a wide range of outcomes, depending on the target of blame (e.g., technical quality, speed, or others). This model is a versatile piece of knowledge to be stored in the designer's toolkit and used whenever needed.

⁴ As pointed out by Mayntz (2004), “it is entirely legitimate to label a mechanism that has been spelled out in detail by a noun that refers to a process, an outcome, or a factor. But to use a terminological label merely to allude to a process that remains unspecified has no more explanatory value than the simple statement of a correlation” (Mayntz 2004: 239).

A) Black-boxed relationship



B) A model of the ICT procedure



C) An abstract model of horizontal monitoring

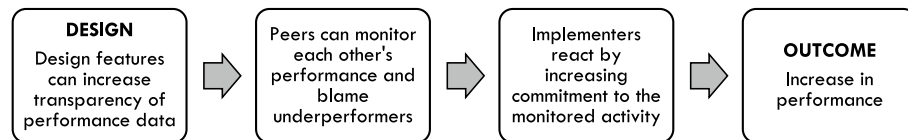


Fig. 2 From the black box to an abstract causal model

Assessing the target context

When using a causal model, the designer faces a typical problem of extrapolation: inferring whether that model can be reproduced in the target context (Steel, 2008). This assessment is needed when a policy program is replicated or scaled up, when the designer is crafting a new intervention, and also when—in light of a change in the context—results deteriorate, and an existing program needs maintenance.

An intuitive way to conduct this assessment is by analyzing the possible analogies and discrepancies between the model and features of the practitioner’s context: whether discrepancies are causally relevant, if they can be disregarded, and if not, whether they can be remedied. To illustrate, Williams (2020) suggests “mechanism mapping” as a graphic diagnostic tool for predicting the impact of scaling up or transporting a policy. After having traced the theory of change of the original program, mechanism mapping requires comparing the contextual assumptions supporting the steps of the theory of change with contextual conditions present in the target context; differences would suggest that the original program will not have the original impact. According to Williams’ analysis, in the case of the Tamil Nadu nutrition program, the step by which “mothers decide to use extra food for selves and infants” is supported by a contextual condition, i.e., “mothers’ control of household food allocation,” which is not present in the Bangladeshi context (Williams, 2020).

These kinds of comparisons are useful but present some fundamental limits. For one thing, the new context may well have the same contextual conditions highlighted in the model but nonetheless vary in some previously unknown features that impair the outcome.

Educated mothers may control household food allocation, for instance, but live in a place where the traditional food culture is so strong that social pressures make them resistant to implementing dietary changes. More generally, the impossibility of knowing all adverse contextual features in advance and the fact that the conditions that one may include as necessary are limited only by a person's imagination (Roberts, 1996) limit the diagnostic utility of these comparisons. Notice, also, that in Williams's (2020) analysis, "household food control" is a contextual condition derived *ex post* from the contrast between the Tamil Nadu and Bangladeshi contexts, but it is doubtful that only by looking at Tamil Nadu, the researcher would have specifically spotted this condition.

Although other authors have proposed methods to estimate the prospects of successful extrapolations (Bates & Glennerster, 2017; Woolcock, 2022), in-depth analysis and context-specific reasoning are required to evaluate the specificities of the practitioner's situation. Causal mechanisms provide a unique piece of information to guide this process. The designer should start with causal powers, identify how they can be exercised in the target context, and analyze whether there are any obstacles to the exercise of these powers. In the example of the nutrition program, for instance, the designer might first identify those who, once educated, could have the power to change the children's diet. Starting with this piece of information avoids the mistake of mechanically including mothers in the program (and involving fathers, mothers-in-law, or whoever possesses that power in this context). Then, the designer investigates if there are any obstacles to the exercise of that power (whether in the household, the general culture of the region, or elsewhere). Past cases and research can help identify relevant contextual features but are no substitute for context-specific assessments.

Designing programs in new contexts

Mechanisms can avoid the use of standard designs across contexts by varying the program in order to fit the target context, conserve the mechanism, and produce the outcome of interest. The result of the contextual assessment described in the previous section may be that no special adjustments are needed. In using model C) in Fig. 2, the designer need only arrange a program to disclose performance information.

If the context is problematic, the designer must work with the elements of the model and arrange a design mix that ensures the reproduction of the modeled mechanism. As discussed in the previous section, this entails the elaboration of functional equivalents, i.e., alternative or additional design features aimed at maintaining or reinforcing the mechanism. The designer starts with the problematic element, its underlying causal power, and devises an appropriate design solution. If peers are resistant to monitoring each other, for instance, the designer may add design features targeting peers more directly (e.g., the nudge feature mentioned earlier). If, working with a different model, time availability is a contextual element that has the power to develop trust but time is scarce in the practitioner's context, the designer can devise certain institutional arrangements that, for instance, increase contacts, facilitate information sharing, and speed up the effects time naturally has on the partners' trust. Similarly, if traditional dietary customs are strong and might pose a social barrier to educating children, the designer may think of additional design features—for instance, including a community-based intervention that can counteract social pressure on mothers.

This fitting process of mechanisms, design, and context is obviously also important when replicating the same program across contexts. If the original design is infeasible, the designer must devise a new design mix that works as a functional equivalent. In replicating the ICT program for a task where digitization is impossible, the designer needs to come up with a functional equivalent possessing the same causal power of increasing transparency, triggering peer monitoring and blame avoidance, and ultimately, increasing commitment and speed.

Conclusion

This article discusses how to investigate causal mechanisms to inform practice, what practical advantages this knowledge may offer, and how to use mechanisms when replicating, adjusting, and designing policy programs. This concluding section offers three additional comments on why this research agenda is worth pursuing.

First, causal mechanisms give practitioners synthetic advice, which is a critical feature of usable knowledge. The too-many-variable problem is well-known in the social sciences. Practitioners can be easily overloaded if they attempt to retrieve and interpret all the potential success factors identified in the literature. Models of causal mechanisms do not include all-encompassing lists of variables nor all the tiny details of an intervention. On the one hand, they do not consider all the classes of critical variables that may affect how a program could work in general but rather only a selection of elements relevant to the outcome of interest; they are not general frameworks but “partial models.” On the other, they simplify the analysis of the program by abstracting causal powers that may be shared by many program and non-program elements. Further, causal powers shed light on the details of actual programs by discriminating between those that are unimportant and contingent on a case and those that are causally relevant to the effectiveness of the intervention.

Second, mechanisms do not provide an instruction sheet for easily “assembling” an intervention. They work like a compass that requires reading and interpretation. By abstracting causal powers, the model identifies the role that design, implementation, and context features are assumed to play. Then, it is up to practitioners to use their knowledge, experience, and ability to read their context and devise a context-specific intervention. This research agenda envisages a potential division of labor between scholars and practitioners, with the former reverse-engineering policy programs and elaborating usable causal models and the latter arranging context-specific applications of those abstract models. Significantly, this points to a clear role for scholars who want to inform practice, namely, enlarging and diversifying the existing knowledge of causal mechanisms of policy programs.

Finally, models of mechanisms provide a kind of knowledge that is flexible across contexts but can support practitioners with design and implementation problems *in specific times and places*. In this respect, studying mechanisms avoids two dangers looming for public policy and administration scholarship. On the one hand, there is the risk of mimicking the natural sciences by searching exclusively for general relationships that are invalid in specific contexts and hardly adjustable to them. This knowledge can help understand what works, for whom, and where, but it requires mechanisms for providing context-specific advice. On the other hand, there is the risk of elaborating hyper-contextualized, one-case descriptions or explanations that, although rich and informative, are impossible to use outside the original context and cumulate as general design knowledge. Causal mechanisms

offer a middle ground worth exploring for the sake of increased practical relevance in public policy and administration.

Acknowledgments I am in debt to Robert Ackrill, Giliberto Capano, the late Bruno Dente, Erica Melloni, Leslie Pal, and Giancarlo Vecchi for reading earlier drafts and providing helpful comments. I also thank the three anonymous reviewers for their constructive feedback.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Ackrill, R., & Kay, A. (2006). Historical-institutionalist perspectives on the development of the EU budget system. *Journal of European Public Policy*, 13(1), 113–133. <https://doi.org/10.1080/13501760500380775>
- Al-Ubaydli, O., List, J., & Suskind, D. (2019). The science of using science: towards an understanding of the threats to scaling experiments. *National Bureau of Economic Research*. <https://doi.org/10.3386/w25848>
- Bailey, M. T. (2006). Do physicists use case studies? Thoughts on public administration research. *Public Administration Review*, 52(1), 47. <https://doi.org/10.2307/976545>
- Bardach, E. (2004). Presidential address? The extrapolation problem: How can we learn from the experience of others? *Journal of Policy Analysis and Management*, 23(2), 205–220. <https://doi.org/10.1002/pam.20000>
- Barzelay, M. (2007). Learning from second-hand experience: Methodology for extrapolation- oriented case research. *Governance*, 20(3), 521–543. <https://doi.org/10.1111/j.1468-0491.2007.00369.x>
- Barzelay, M. (2019). *Public management as a design-oriented professional discipline*. Edward Elgar.
- Bates, M. A., & Glennerster, R. (2017). The generalizability puzzle. *Stanford Social Innovation Review*, 15(3), 50–54.
- Beach, D., & Pedersen, R. B. (2016). *Causal case study methods*. University of Michigan Press.
- Beach, D., & Pedersen, R. B. (2018). Selecting appropriate cases when tracing causal mechanisms. *Sociological Methods and Research*, 47(4), 837–871. <https://doi.org/10.1177/0049124115622510>
- Bennett, A. (2013). The mother of all isms: Causal mechanisms and structured pluralism in international relations theory. *European Journal of International Relations*, 19(3), 459–481. <https://doi.org/10.1177/1354066113495484>
- Betts, J. R., & Shkolnik, J. L. (1999). The behavioral effects of variations in class size: The case of math teachers. *Educational Evaluation and Policy Analysis*, 21(2), 193–213. <https://doi.org/10.3102/01623737021002193>
- Bhaskar, R. (2013). *A realist theory of science*. Routledge.
- Bobrow, D. B. (2006). Policy design: Ubiquitous necessary and difficult. In B. G. Peters & J. Pierre (Eds.), *Handbook of public policy* (pp. 75–96). Sage.
- Bourke, S. (1986). How smaller is better: Some relationships between class size, teaching practices, and student achievement. *American Educational Research Journal*, 23(4), 558–571. <https://doi.org/10.3102/00028312023004558>
- Bradley, E. H., Curry, L. A., Ramanadhan, S., Rowe, L., Nembhard, I. M., & Krumholz, H. M. (2009). Research in action: Using positive deviance to improve quality of health care. *Implementation Science*, 4(1), 1–11. <https://doi.org/10.1186/1748-5908-4-25>
- Brunner, R. D. (2006). A paradigm for practice. *Policy Sciences*, 39(2), 135–167. <https://doi.org/10.1007/s11077-006-9012-9>
- Bunge, M. (2004). How does it work?: The search for explanatory mechanisms. *Philosophy of the Social Sciences*, 34(2), 182–210. <https://doi.org/10.1177/0048393103262550>

- Busetti, S., & Dente, B. (2021). When red tape saves time: The anti-corruption controls for the 2015 universal exposition. *International Review of Public Policy*, 3(1), 1.
- Busetti, S., & Dente, B. (2018). Designing multi-actor implementation: A mechanism-based approach. *Public Policy and Administration*, 33(1), 46–65. <https://doi.org/10.1177/0952076716681207>
- Cammett, M. (2022). Positive deviance cases: Their value for development research policy and practice. In J. Widner, M. Woolcock, & D. Ortega Nieto (Eds.), *The Case for Case Studies: Methods and Applications in International Development* (pp. 219–238). Cambridge University Press.
- Capano, G., & Howlett, M. (2019). Causal logics and mechanisms in policy design: How and why adopting a mechanistic perspective can improve policy design. *Public Policy and Administration*. <https://doi.org/10.1177/0952076719827068>
- Capano, G. (2019). Reconceptualizing layering—from mode of institutional change to mode of institutional design: Types and outputs. *Public Administration*, 97(3), 590–604. <https://doi.org/10.1111/padm.12583>
- Cartwright, N. (2007). *Hunting causes and using them: Approaches in philosophy and economics*. Cambridge University Press.
- Cartwright, N., & Hardie, J. (2012). *Evidence-based policy: A practical guide to doing it better*. Oxford: Oxford University Press.
- Chikofsky, E. J., & Cross, J. H. (1990). Reverse engineering and design recovery: A taxonomy. *IEEE Software*, 7(1), 13–17. <https://doi.org/10.1109/52.43044>
- Collier, A. (1994). *Critical Realism: An Introduction to Roy Bhaskar's Philosophy*. Verso. <https://philpapers.org/rec/COLCRA>
- Davies, H. T. O., Nutley, S. M., & Smith, P. C. (Eds.). (2000). *What works?: Evidence-based policy and practice in public services*. The Policy Press.
- DeLeon, P. (1988). The contextual burdens of policy design. *Policy Studies Journal*, 17(2), 297–309.
- Dryzek, J. S., & Ripley, B. (1988). The ambitions of policy design. *Review of Policy Research*, 7(4), 705–719. <https://doi.org/10.1111/j.1541-1338.1988.tb00890.x>
- Durant, R. F. (1984). EPA, TVA and pollution control: Implications for a theory of regulatory policy implementation. *Public Administration Review*, 44(4), 315. <https://doi.org/10.2307/976076>
- Elster, J. (1989). *Nuts and bolts for the social sciences*. Cambridge University Press.
- Falletti, T. G., & Lynch, J. F. (2009). Context and causal mechanisms in political analysis. *Comparative Political Studies*, 42(9), 1143–1166. <https://doi.org/10.1177/0010414009331724>
- Finn, J. D., & Achilles, C. M. (1990). Answers and questions about class size: A statewide experiment. *American Educational Research Journal*, 27(3), 557–577. <https://doi.org/10.3102/00028312027003557>
- Finn, J. D., & Achilles, C. M. (1999). Tennessee's class size study: Findings, implications, misconceptions. *Educational Evaluation and Policy Analysis*, 21(2), 97–109. <https://doi.org/10.3102/01623737021002097>
- Finn, J. D., Pannozzo, G. M., & Achilles, C. M. (2003). The “Why’s” of class size: Student behavior in small classes. *Review of Educational Research*, 73(3), 321–368. <https://doi.org/10.3102/00346543073003321>
- Flyvbjerg, B. (2001). *Making social science matter*. Cambridge University Press.
- Funnell, S. C., & Rogers, P. J. (2011). *Purposeful program theory*. Jossey-Bass.
- Gero, J. S. (1990). Design prototypes: A knowledge representation schema for design. *AI Magazine*, 11(4), 36. <https://doi.org/10.1609/AIMAG.V11I4.854>
- Gerring, J. (2010). Causal mechanisms: Yes, but. *Comparative Political Studies*, 43(11), 1499–1526. <https://doi.org/10.1177/0010414010376911>
- Gerring, J. (2007). *Case study research: Principles and practices*. Principles and Practices. Cambridge University Press. <https://doi.org/10.1017/CBO9780511803123>
- Gerring, J. (2008). The mechanistic worldview: Thinking inside the box. *British Journal of Political Science*, 38(1), 161–179. <https://doi.org/10.1017/S0007123408000082>
- Gill, J., & Meier, K. J. (2001). Ralph's pretty-good grocery versus Ralph's super market: Separating excellent agencies from the good ones. *Public Administration Review*, 61(1), 9–17. <https://doi.org/10.1111/0033-3352.00002>
- Graue, E., Hatch, K., Rao, K., & Oen, D. (2007). The wisdom of class-size reduction. *American Educational Research Journal*, 44(3), 670–700. <https://doi.org/10.3102/0002831207306755>
- Green, D. P., Ha, S. E., & Bullock, J. G. (2010). Enough already about “black box” experiments: Studying mediation is more difficult than most scholars suppose. *Annals of the American Academy of Political and Social Science*, 628(1), 200–208. <https://doi.org/10.1177/0002716209351526>
- Greenland, S., Pearl, J., & Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology*, 10(1), 37–48.

- Hedström, P., & Ylikoski, P. (2010). Causal mechanisms in the social sciences. *Annual Review of Sociology*, 36, 49–67. <https://doi.org/10.1146/annurev.soc.012809.102632>
- Hogwood, B. W., & Gunn, L. A. (1984). *Policy analysis for the real world*. Oxford University Press.
- Kay, A. (2005). A critique of the use of path dependency in policy studies. *Public Administration*, 83(3), 553–571. <https://doi.org/10.1111/j.0033-3298.2005.00462.x>
- Lasswell, H. D., & Kaplan, A. (1950). *Power and society: A framework for political inquiry*. Yale University Press.
- Lasswell, H. D. (1970). The emerging conception of the policy sciences. *Policy Sciences*, 1(1), 3–14. <https://doi.org/10.1007/BF00145189>
- Leach, W. D., & Pelkey, N. W. (2001). Making watershed partnerships work: A review of the empirical literature. *Journal of Water Resources Planning and Management*, 127(6), 378–385. [https://doi.org/10.1061/\(ASCE\)0733-9496\(2001\)127:6\(378\)](https://doi.org/10.1061/(ASCE)0733-9496(2001)127:6(378))
- Levin, J. (2002). For whom the reductions count: A quantile regression analysis of class size and peer effects on scholastic achievement. In Fitzenberger B., Koenker R., & Machado J.A.F. (Eds.), *Economic Applications of Quantile Regression* (pp. 221–246). Physica. https://doi.org/10.1007/978-3-662-11592-3_11
- Lindblom, C. E., & Cohen, D. K. (1979). *Usable knowledge*. Yale University Press. <https://doi.org/10.1109/TSMC.1980.4308493>
- Lindquist, E. A., Howlett, M., Skogstad, G., Tellier, G., & 't Hart, P. (2022). *Policy success in Canada: Cases, lessons*. Oxford University Press.
- Luetjens, J., Mintrom, M., & 't Hart, P. (2019). *Successful public policy. Lessons from Australia and New Zealand*. ANU Press. <https://doi.org/10.22459/SPP.2019>
- Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science*, 67(1), 1–25. <https://doi.org/10.1086/392759>
- Mahoney, J. (2001). Beyond correlational analysis: Recent innovations in theory and method. *Sociological Forum*, 16(3), 575–593.
- May, P. J. (1992). Policy learning and failure. *Journal of Public Policy*, 12(4), 331–354.
- Mayne, J. (2012). Contribution analysis: Coming of age? *Evaluation*, 18(3), 270–280. <https://doi.org/10.1177/1356389012451663>
- Mayntz, R. (2004). Mechanisms in the analysis of social macro-phenomena. *Philosophy of the Social Sciences*, 34(2), 237–259. <https://doi.org/10.1177/0048393103262552>
- McAdam, D., Tarrow, S., & Tilly, C. (2004). *Dynamics of contention*. Cambridge University Press.
- McConnell, A. (2010). *Understanding policy success: Rethinking public policy*. Palgrave Macmillan.
- Meier, K. J., & Gill, J. (2000). *What works: A new approach to program and policy analysis*. Westview Press.
- Melloni, E. (2013). Ten years of European impact assessment world political science review 2013; aop ten years of European impact assessment: how it works, for what and for whom. *World Political Science*, 9(1), 263–290. <https://doi.org/10.1515/wpsr-2013-0011>
- Melloni, E., Pesce, F., & Vasilescu, C. (2016). Are social mechanisms usable and useful in evaluation research? *Evaluation*, 22(2), 209–227. <https://doi.org/10.1177/1356389016643900>
- Missildine, K., Fountain, R., Summers, L., & Gosselin, K. (2013). Flipping the classroom to improve student performance and satisfaction. *Journal of Nursing Education*, 52(10), 597–599. <https://doi.org/10.3928/01484834-20130919-03>
- O'Toole, L. J. (1986). Policy recommendations for multi-actor implementation: An assessment of the field. *Journal of Public Policy*, 6(2), 181–210. <https://doi.org/10.1017/S0143814X00006486>
- O'Toole, L. J. (2000). Research on policy implementation: Assessment and prospects. *Journal of Public Administration Research and Theory*, 10(2), 263–288. <https://doi.org/10.1093/oxfordjournals.jpart.a024270>
- Ongaro, E. (2009). *A protocol for the extrapolation of 'Best' Practices: How to draw lessons from one experience to improve public management in another situation*. European Public Sector Award 2009, Final Symposium and Ceremony, Maastricht. Available at: http://epsa2009.eu/files/Symposium/An%20approach%20to%20the%20extrapolation%20of%20practices_EOngaro.pdf
- Overman, E. S., & Boyd, K. J. (1994). best practice research and postbureaucratic reform. *Journal of Public Administration Research and Theory*, 4(1), 67–83. <https://doi.org/10.1093/oxfordjournals.jpart.a037195>
- Pawson, R., & Tilley, N. (1997). *Realistic evaluation*. Sage Publications.
- Pawson, R. (2006). *Evidence-based policy: A realist perspective*. USA: Sage Publications.
- Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4), 669–688. <https://doi.org/10.1093/BIOMET/82.4.669>
- Pearl, J., & Mackenzie, D. (2018). *The book of why*. Basic Books.

- Pielke, R. A. (2004). What future for the policy sciences? *Policy Sciences*, 37(3–4), 209–225. <https://doi.org/10.1007/s11077-005-6181-x>
- Raadschelders, J. C. (2011). The future of the study of public administration: Embedding research object and methodology in epistemology and ontology. *Public Administration Review*, 71(6), 916–924.
- Ragin, C. C. (2008). *Redesigning social inquiry: Fuzzy sets and beyond*. The University of Chicago Press.
- Roberts, C. (1996). *Logic of historical explanation*. The Pennsylvania State University Press.
- Rose, R. (1993). *Lesson drawing in public policy: A guide to learning across time and space*. Chatham House.
- Rueschemeyer, D. (2009). *Usable theory: Analytic tools for social and political research*. Princeton University Press.
- Sabatier, P., & Mazmanian, D. (1980). The implementation of public policy: A framework of analysis. *Policy Studies Journal*, 8(4), 538–560. <https://doi.org/10.1111/j.1541-0072.1980.tb01266.x>
- Sayer, A. (1992). *Method in social science: A realist approach*. USA: Routledge.
- Schön, D. A. (1983). *The reflective practitioner*. Basic Books.
- Seawright, J. (2016). The case for selecting cases that are deviant or extreme on the independent variable. *Sociological Methods and Research*, 45(3), 493–525. <https://doi.org/10.1177/0049124116643556>
- Stake, R. E. (2010). *Qualitative research: Studying how things work*. The Guilford Press.
- Steel, D. (2004). Social Mechanisms and Causal Inference. *Philosophy of the Social Sciences*, 34(1), 55–78. <https://doi.org/10.1177/0048393103260775>
- Steel, D. P. (2008). *Across the boundaries: Extrapolation in biology and social science*. Oxford: Oxford University Press.
- Tjalve, E. (1979). *A Short Course in Industrial Design*. Butterworth & Co.
- van der Heijden, J., Kuhlmann, J., Lindquist, E., & Wellstead, A. (2019). Have policy process scholars embraced causal mechanisms? *Public Policy and Administration*. <https://doi.org/10.1177/0952076718814894>
- Van Meter, D. S., & Van Horn, C. E. (1975). The policy implementation process: A conceptual framework. *Administration & Society*, 6(4), 445–488. <https://doi.org/10.1177/009539977500600404>
- Waldo, D. (1955). *The study of public administration*. Doubleday.
- Weaver, R. K. (2019). Reverse engineering and policy design. In G. Capano, M. Howlett, M. Ramesh, & A. Virani (Eds.), *Making Policies Work* (pp. 173–190). Edward Elgar Publishing. <https://doi.org/10.4337/9781788118194.00020>
- Weiss, C. H. (1997). Theory-based evaluation: Past, present, and future. *New Directions for Evaluation*, 1997(76), 41–55. <https://doi.org/10.1002/EV.1086>
- Williams, M. J. (2020). External validity and policy adaptation: From impact evaluation to policy design. *The World Bank Research Observer*, 35(2), 158–191. <https://doi.org/10.1093/WBRO/LKY010>
- Woolcock, M. (2022). Will It Work Here? Using Case Studies to Generate ‘Key Facts’ About Complex Development Programs. In J. Widner, M. Woolcock, & D. Ortega Nieto (Eds.), *The Case for Case Studies Methods and Applications in International Development* (pp. 87–115). Cambridge University Press.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.