Systems Approaches to Public Service Delivery:

Methods and Frameworks

August 2022

Zahra Mansoor and Martin J. Williams*

Abstract

Researchers and practitioners are increasingly embracing *systems approaches* to deal with the complexity of public service delivery and policy evaluation. However, the diversity of these methods and their lack of common theoretical grounding has limited constructive engagement between those working within the systems tradition and those working outside it. We address this by reviewing and critically synthesizing systems literature from the fields of health, education, and infrastructure. We argue that the common theoretical core of systems approaches is the idea that *multi-dimensional complementarities* between a policy and other aspects of the policy context are the first-order problem of policy design and evaluation. What differentiates systems approaches from other research traditions is thus not so much a specific method as a general difference in question prioritization, and consequently greater methodological pluralism. We distinguish between *macro-systems* approaches, which focus on the collective coherence of a set of policies or institutions, and *micro-systems* approaches, which focus on how a single policy interacts with the context in which it operates. We develop a typology of micro-systems approaches and their relationship to standard impact evaluation methods, and discuss their relationship to adjacent concepts such as external validity, implementation science, and complexity theory.

* Corresponding author: Zahra Mansoor, Postdoctoral Researcher, Blavatnik School of Government, University of Oxford, zahra.mansoor@bsg.ox.ac.uk. Williams: Associate Professor in Public Management, Blavatnik School of Government, University of Oxford. We are grateful for funding from the Bill and Melinda Gates Foundation and helpful conversations and comments from Seye Abimbola, Dan Berliner, Lucy Gilson, Guy Grossman, Kara Hanson, Rachel Hinton, Dan Honig, Adnan Khan, Julien Labonne, Lant Pritchett, Imran Rasul, Joachim Wehner, and workshop participants at the Blavatnik School of Government and 2018 Global Symposium on Health Systems Research. The authors are responsible for any remaining faults. Declarations of interest: none.

1. Introduction

Across the social sciences, researchers and practitioners working to use evidence to improve public service delivery are increasingly turning to *systems approaches* to remedy what they see as the limitations of traditional approaches to policy evaluation. This includes increasing calls from disciplines like economics and management to adopt systems approaches to understanding the complexities of government bureaucracies (Pritchett, 2015; Bandiera et al., 2019; Besley et al., 2021). While those turning to systems approaches are united in viewing standard impact evaluation methods (at least in their more naïve applications) as overly simplistic, deterministic, and insensitive to context, the alternative methods they have developed are hugely varied. Studies that self-identify as systems approaches include everything from ethnographic approaches to understanding citizen engagement with public health campaigns during the 2014 Ebola outbreak in West Africa (Martineau, 2016) to high-level World Health Organization frameworks (De Savigny and Adam, 2009), multi-sectoral computational models of infrastructure systems (e.g. Saidi et al., 2018), diagnostic surveys to identify system weaknesses (Halsey and Demas, 2013), and "whole-of-government" governance approaches to address the new cross-sectoral coordination challenges imposed by Covid-19 (OECD 2017). This extreme diversity in concepts and methods can make systems approaches seem ill-defined and opaque to researchers and policymakers from outside the systems tradition, and has limited engagement with their insights.

What, then, is the common theoretical core of systems approaches to public service delivery? What are the key distinctions among them, and to which kinds of questions or situations are different types of systems approaches best suited? And what is the relationship between

systems approaches and standard impact evaluation-based approaches to using evidence to improve public service delivery?

We address these questions by reviewing and synthesizing the growing literature on systems approaches. We focus our review on three policy sectors in which systems approaches have gained increasing currency in high- as well as middle- and low-income countries alike: health, education, and infrastructure. These approaches have developed largely independently in each sector, which not only creates opportunities for learning across sectors but also allows us to distill a common set of conceptual underpinnings from a diverse array of methods, contexts, and applications.

Our article thus has two linked goals. First, we aim to provide shared conceptual foundations for engagement between researchers within the systems tradition and those who work outside the systems community but share an interest in the role of context and complexity in public service delivery and policy evaluation. Second, we aim to cross-pollinate ideas and facilitate discussion within the systems research community, among researchers and practitioners from different sectoral backgrounds or disciplinary communities.

Based on our review, we argue that systems approaches can best be understood not as a single method, but as a diverse set of analytical responses to the idea that *multi-dimensional complementarities* between a policy and other aspects of the policy's context (e.g. other policies, institutions, social and economic context, cultural norms, etc.) are the first-order problem of policy design and evaluation. Such complementarities are present when the impact of a group of variables on an outcome is greater than the sum of its parts. For example, the impact of a new pay-for-performance scheme on health service delivery might depend not just

on multiple characteristics of the scheme's design but also on the presence of effective data monitoring and auditing systems, on health workers' intrinsic motivation and career incentives, on the availability of resources to pay bonuses, and on whether political economy considerations permit the non-payment of bonuses – as well as potentially dozens of other dimensions along which contexts might vary. Whereas standard impact evaluation methods typically seek to address these complexities by finding a way to "hold all else constant" in order to causally identify the impact of a policy intervention on an outcome variable, systems approaches focus in on the "all else" in order to better understand the complex ways in which policies' effectiveness might vary across contexts and time or depend on the presence of complementary policy interventions.

Within the broad umbrella of systems approaches, we distinguish between "macro-systems" approaches and "micro-systems" approaches. The former are primarily concerned with understanding the collective coherence of a set of policy interventions and various other elements of context, whereas the latter focus in on a single policy intervention (like most standard impact evaluations) but focus on understanding its interactions with contextual variables and other policy interventions (rather than necessarily obtaining an average treatment effect). We further review and distinguish among different analytical methods within each of these two categories, and link these different methods to different questions and analytical purposes. In particular, we suggest that the choice of which micro-systems approach to adopt depends on the degree to which contextual complementarities affect a policy's *efficacy* (i.e. the extent to which a given policy has consistent impacts across contexts) and *implementability* (i.e the extent to which a given policy can be delivered or implemented correctly). We combine these two dimensions to construct four stylized types of linked question types and research approaches: "what works"-style impact evaluation (consistent efficacy, consistent

implementability); external validity (inconsistent efficacy, consistent implementability); implementation science (consistent efficacy, inconsistent implementability); and complex systems (inconsistent efficacy, inconsistent implementability). While not necessarily straightforward to apply in practice, this parsimonious framework helps explain why and when researchers might choose to adopt different systems-based methods to understand different policies and different questions – as well as when adopting a systems perspective may be less necessary.

Of course, these questions are also of interest to impact evaluators outside the systems tradition, and many of the methodological tools that systems researchers use are familiar to them. Whereas systems approaches are sometimes perceived as being from a different epistemological tradition than standard impact evaluation methods (e.g. Marchal et al., 2012) we view the underlying epistemology of systems approaches as consistent with that of impact evaluation. The main difference is the extent to which complementarities are relevant and hence how tractable understanding their impact is through standard impact evaluation methods with limited statistical power and counterfactual availability. While issues of heterogeneity, complementarity, and external validity can be addressed using standard impact evaluation methods (e.g. Bandiera et al., 2010; Andrabi et al., 2020), systems approaches presume (implicitly or explicitly) that such interactions are often high-dimensional (i.e. across many different variables) and thus intractable with limited sample sizes.[1] What distinguishes systems approaches, then, is mainly a different *prioritization* of these questions, and consequently a greater openness to methods other than quantitative impact evaluation in answering them. In this view, systems approaches and impact evaluation are thus better understood as

[1] See Hausmann (2008), Pritchett (2015), and Williams (2020), among others, for related discussions.

complements, not mutually inconsistent alternatives, for creating and interpreting evidence about policy effectiveness.

The remainder of our article proceeds as follows. Section 2 briefly discusses our review method. Section 3 presents a range of definitions of systems approaches from the literature, then synthesizes them into what we characterize as their common theoretical core. Section 4 reviews and typologizes macro-systems approaches across health, education, and infrastructure and offers a conceptual framework for synthesis, and Section 5 does the same for micro-systems approaches. Section 6 discusses how researchers and practitioners should go about selecting which type of systems approach (if any) is best suited for their purposes, and Section 7 concludes by discussing the connections between systems approaches to public service delivery and other well-established theoretical and methodological concerns in economics, political science, and public administration.

2. Review Method

Our review of systems approaches in public service delivery focuses primarily on three sectors in which they have increasingly gained popularity: health, education, and infrastructure. However, the purpose of this article is not to provide a comprehensive survey of the systems literature in each of these sectors, as there already exist several excellent sector-level survey papers on systems approaches (e.g. Gilson, 2012; Carey et al., 2015; Hanson, 2015 for health; Pritchett, 2015 for education; Saidi et al., 2018 for infrastructure). Instead, this article's main contribution is to synthesize ideas and insights from these divergent sectoral literatures to make them more accessible to each other and to readers from outside the systems tradition.

Given this, we opted not to use a formal systematic review methodology but rather to conduct our review through a combination of citation-tracing from foundational texts and keyword searches in online databases, supplemented by suggestions from sectoral experts. We conducted selective literature reviews within each sector aimed at synthesizing the breadth of questions, theories, methods, and empirical applications that comprise the range of methods used in the systems literature across these sectors. We also draw on non-sector-specific work on systems approaches to understanding service delivery in complex and unpredictable systems more generally. We include in our review texts that self-describe as systems-based, as well as many which share similar concerns and methods but which do not necessarily adopt the language of systems approaches.

For clarity and brevity, and in line with the article's purpose, we focus the main text on presenting an overall synthesis with illustrative examples rather than on decreasing readability by trying to cover as many studies as possible. We include a more detailed (though still inevitably selective) sector-by-sector summary in an Online Appendix for interested readers.

Our review and synthesis is not necessarily intended as an argument in favor of systems approaches being used more widely, nor as a critique of research outside the systems tradition. Neither should it be read as a critique of systems approaches. While we do believe that both the general thrust of systems approaches and many of the specific ideas presented by them are important and useful, our goal is merely to present a concise survey and a set of clear conceptual distinctions so that readers can determine what might be useful to them from within this diverse array of perspectives and methods and can better converse across disciplinary and sectoral boundaries without the caricaturing and misrepresentation that have often marred these conversations. Doing this inevitably creates a tension between staying faithful to the way in

which researchers in these fields view their work, and the necessity of communicating about it in ways that will be intelligible to readers from other fields. We hope that we have struck this balance well and that readers will be understanding of the challenges of doing so on such a broad-ranging topic.

3. Defining Systems Approaches

Systems approaches are defined in different ways across different sectors, but tend to share a common emphasis on the multiplicity of actors, institutions, and processes within systems. For example, the World Health Organization (2007, p. 2) defines a health system as consisting of "all organizations, people and actions whose primary intent is to promote, restore or maintain health." In education, Moore (2015, p. 1) defines education systems as "institutions, actions and processes that affect the 'educational status' of citizens in the short and long run." In infrastructure, Hall et al. (2016, p. 6) define it as "the collection and interconnection of all physical facilities and human systems that are operated in a coordinated way to provide a particular infrastructure service".

Despite their differences, these definitions imply a focus of systems on "holism" (Midgley 2006; Hanson 2015), or the idea that individual policies do not operate in isolation. Whereas a great deal of research and evidence-based policymaking focuses on studying the effectiveness of a single policy in isolation – often by means of using impact evaluation to estimate an average treatment effect – in practice each policy's effectiveness depends on other policies and various features of the contextual environment (Hanson, 2015). As De Savigny and Adam (2009, p.19) write in their seminal discussion of health systems, "every intervention, from the simplest to the most complex, has an effect on the overall system, and the overall system has

an effect on every intervention." This emphasis on interconnection has made the study of complexity (e.g. Stacey, 2010; Burns & Worsley, 2015) a natural source of inspiration for those seeking to apply systems approaches to the study of development and public service delivery.

But despite the growing popularity of systems approaches, there remains significant ambiguity around their meaning, with no universally accepted definition or conceptual framework beyond their shared emphasis on holism, context, and complexity (Midgley, 2006). Even those writing within the systems tradition have pointed out that the field has used "diverse" and "divergent" concepts and definitions, leading the field as a whole to be sometimes characterized as "ambiguous" and "amorphous" (Cabrera et al., 2008). This lack of a commonly agreed definition and theoretical basis has made a precise and concise response to the question "what is a systems approach to public service delivery, and how is it different to what already exists?" difficult to obtain.

We argue that instead of viewing a systems approach as a specific method, system approaches are better understood as a diverse set of analytical responses to the idea that the first-order challenge of policy design and evaluation is to understand the multi-dimensional complementarities between a policy and other aspects of the policy's context (e.g. other policies, institutions, social and economic context, cultural norms, etc.). By complementarities, we refer to the formal definition under which two variables – e.g. a variable capturing the presence of a particular policy and another variable capturing some aspect of the policy's context – are considered complements when their joint effect on an outcome variable is greater than the sum of their individual effects on that variable.[2] By multi-dimensional, we refer to the

[2] The prevalence of complementarities in bureaucracies has also been emphasized in organizational research (e.g. Ichniowski and Shaw 2003; Brynjolfsson and Milgrom 2013) and used in explaining institutional path dependence (Deeg, 2007).

idea that these complementarities might not just be among two or three variables at a time (as impact evaluations often seek to estimate) but among so many variables that estimating them in a standard econometric framework often becomes intractable. While this definition is limited in its precision by the need to adequately encompass the enormous diversity of systems approaches we discuss in subsequent sections, it captures the theoretical core – the emphasis on understanding multi-dimensional complementarities – that ties them all together.

Advocates of systems approaches often contrast this emphasis with the naïve use of impact evaluation to obtain an average treatment effect of a policy which is then used to guide adoption decisions across a wide range of contexts and populations. Of course, the rapid growth in attention towards and research on issues of external validity and implementation within economics and political science (Deaton, 2010; Pritchett & Sandefur, 2015; Bold et al., 2018) makes this something of a "straw-man" characterization in many cases. In practice, both "impact evaluators" and "systems researchers" care about average treatment effects as well as about heterogeneity, mechanisms, and interactions. But while easily over-exaggerated, the distinction does capture the different frame of mind with which systems researchers approach evidence-based policy, in which understanding complementarities among policies and their context is the *primary* focus of analysis, prioritized (in many cases) even over estimating the direct effect of a policy itself. Whereas a standard impact evaluation seeks primarily to understand the impact of a specific policy holding all else constant, a systems approach to the same policy seeks primarily to understand how the "all else" affects the policy's impacts.

Among studies that self-identify as focusing on systems, one can draw a conceptual distinction between studies that are system-focused *in substance* (due to their scale or topic) and those that are system-focused *in approach* (due to their methodological or theoretical emphasis on issues

of context, complementarity, and contingency). This article focuses mainly on the latter category. Although in practice these categories overlap significantly and the distinction is a blurry one, it nonetheless helps avoid the excessive conceptual spread that could result from referring to every study on "the health system" (or the education or infrastructure systems) as a "systems approach".

Before we proceed to drawing distinctions among different types of systems approaches, it is worth noting two additional characterizations of systems approaches that are often made by systems researchers. First, systems approaches are sometimes viewed as being more question- or problem-driven than standard research approaches, with a focus on real world issues and linkages to actual government policy choices (e.g. Mills, 2012; Gilson, 2012; Hanson, 2015). While this characterization risks giving short shrift to the policy relevance of a great deal of research outside the systems tradition, there is also a natural linkage between embeddedness in an actual policy decision and a concern for understanding how a wide range of factors interlock, since policymakers must often deal with a breadth of challenges that researchers might choose to abstract away in the pursuit of parsimony. Second, some systems researchers emphasize that service delivery is not only complicated (in the sense of involving many moving parts) but also complex (in the sense of possessing dynamics that are non-linear and/or fundamentally unpredictable) (Sheikh et al., 2011, Snyder, 2013). We do not include this aspect of complexity in our core definition presented above, since it is far from universally shared among systems approaches, but return to discuss this issue further in section 6 below.

4. Macro-systems Approaches

One branch of systems approaches responds to the challenge posed by the presence of multi-dimensional complementarities across policies and contextual factors by taking a step back to try to examine questions of policy effectiveness from the standpoint of the entire system. These *macro-systems approaches* are focused not on the impact of a specific policy in isolation, but on understanding how the entire system functions to deliver desired outcomes. Macro-systems approaches thus focus on understanding coherence and interconnectedness between different policies, structures, and processes. In doing so, they also tend to define boundaries of the system in question, although this is often a challenging task (Carey et al., 2015).

Our review of macro-systems approaches across the health, education, and infrastructure sectors highlights that these approaches lie on a spectrum of the specificity with which they define causal relationships between different system components. This includes approaches ranging from those that merely outline lists or typologies of various system components to those that tend to specify causal relationships between system components through specific numerical parameters. Along this spectrum it is possible to distinguish three types of macro-systems approaches:

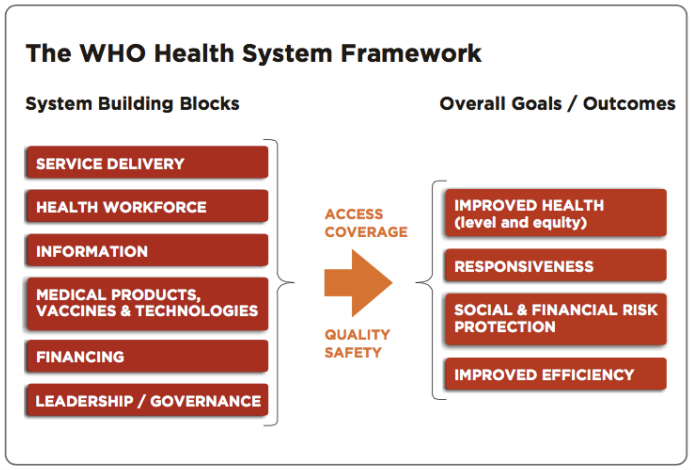
- *Inventory approaches,* which are primarily descriptive and use typologies or lists to define a comprehensive universe of system components such as the types of stakeholders, functions, institutions, or processes within a system;

- *Relational approaches,* which go a step further to posit broad causal relationships or complementarities between system features, based mainly on theory[3]; and
- *Systems modelling,* which conceptualizes the system through precise mathematical causal relationships between different system components.

Inventory approaches list different components and /or typologies within a system with the aim of cataloguing the whole range of factors that determine the outcomes or performance of a given system (usually defined sectorally). An example of such an approach is the seminal WHO health systems framework which characterizes the health system as comprising six key functional building blocks – service delivery, health workforce, information, medical products (including both vaccines and technologies), financing, and leadership and governance – and links them to the broader health system goals (World Health Organization, 2007). As Figure 1 shows, the strength of such inventory frameworks is their very wide scope in terms of identifying the full range of potential determinants and outcomes of a system, but this breadth is achieved by limiting the specificity of the causal relationships they posit. Similarly, the World Bank Systems Approach for Better Education Results (SABER) defines the education system in terms of thirteen different functions (e.g. education management information systems, school autonomy and accountability, student assessment) with a link to improved student learning without specifying the relationship between these functions (Halsey and Demas, 2013).

[3] The "inventory" and "relational" terms are drawn from Hanson's (2015) excellent review of the health systems literature.

Figure 1: WHO Health System Framework

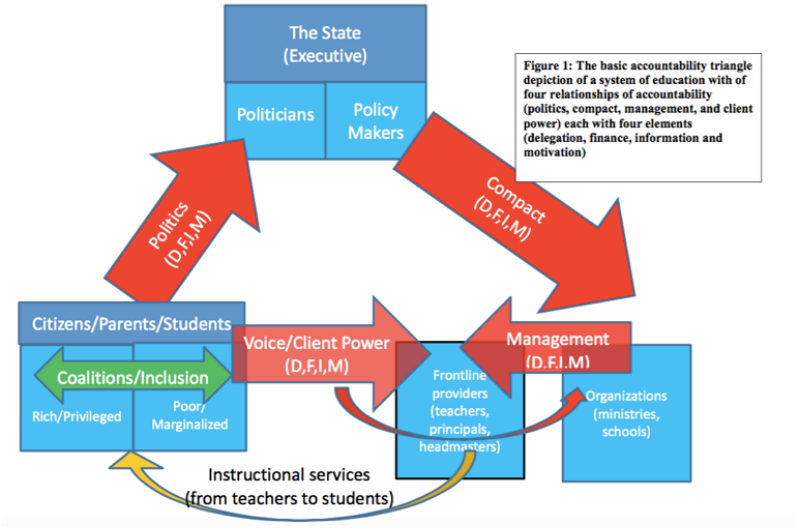


Source: De Savigny and Adam (2009)

Like inventory approaches, relational macro-systems approaches list different system components, but go a step further in specifying the nature or direction of specific relationships or complementarities between them. For example, Gilson (2003) conceptualizes the health system as a set of trust relationships between patients, providers, and the wider institutions. This differs from an inventory approach in more narrowly specifying both the content and direction of relationships among actors, which makes it more analytical but also limits its scope. It also demonstrates how such frameworks may also consider the software (i.e. institutional environment, values, culture and norms) in addition to the hardware (i.e. population, providers, organizations) of a health system (Sheikh et al., 2011). In the education sector, Pritchett (2015) adopts a relational approach to characterizing the education system through accountability links between different actors such as the executive apparatus of the state, organizational providers of schooling (such as ministries and schools), frontline providers

(such as head teachers and teachers), and citizens (such as parents and students).[4] He argues that the system of education works when there is an adequate flow of accountability across the key actors in the system across four design elements: delegation, financing, information and motivation (see Figure 2). Similarly, in the infrastructure sector, Ottens et al. (2006) propose a high-level framework to characterize how technical elements in an infrastructure system may interact with human actors and social institutions to determine system performance. But while such relational approaches are more specific than inventory approaches in their definition of elements and causal relationships, they are still broad enough that their use is more as a conceptual framework for arraying factors and nesting hypotheses than as an operationalizable model of the system.

Figure 2: Education System Framework



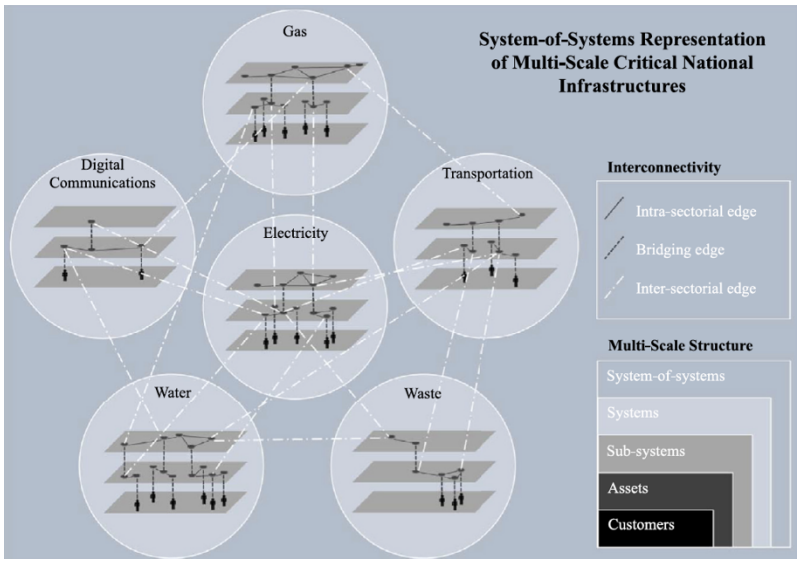
Source: Pritchett (2015)

[4] Pritchett's (2015) framework builds upon the World Bank's (2004) well-known "accountability triangle", itself a relational framework.

Systems modelling approaches take this next step of precisely specifying variables, causal relationships among these system components, and numerical parameters on these relationships. Such models typically combine theory with statistical methods, and draw on a range of quantitative techniques such as systems dynamics, structural equation modelling, and structural econometric modelling (e.g. Reiss & Wolak, 2007; Homer & Hirsch, 2006).[5] Thacker et al. (2017), for example, develop a network-based *systems-of-systems* model for critical national infrastructures, where each type of infrastructure such as water or electricity is a sub-system comprising of a group of nodes and edges with their specific flows (see Figure 3). They use this model to perform a multi-scale disruption analysis and draw predictions on how failures in any individual sub-systems can potentially lead to large disruptions. In the health sector, Homer and Hirsch (2006) develop a causal diagram of how chronic disease prevention works and then use systems dynamic methodology to develop a computer-based model to test alternate policy scenarios that may affect the chronic disease population. In the education sector, Kaffenberger and Pritchett (2021) combine a structural model with parameter values from existing empirical literature to predict how learning outcomes would be affected under different policy scenarios such as expanding schooling to universal basic education, slowing the pace of curriculum, and increasing instructional quality.

[5] Systems dynamics methodology involves computer simulation models to capture processes of accumulation and feedback using numerical values (Homer & Hirsch, 2006). This is related methodologically to the type of formal theoretical and empirical structural modelling methods often used in the social sciences; the distinction between them lies less in the methods themselves than in the intent to model relationships across an entire system or sub-system.

Figure 3: Infrastructure System representation with six critical national infrastructures



Source: Thacker et al. (2017)

The three macro-systems approaches outlined above can have different types of uses and benefits depending on the question of interest. For example, systems researchers often use frameworks developed through inventory approaches to develop diagnostic tools to understand strengths and weaknesses of systems. For example, the World Bank has used its SABER framework to develop system diagnostic tools which have been implemented in more than 100 countries to identify key constraints to system effectiveness and the relationship between different system components (World Bank, 2014). Relational frameworks in turn can be used to array key relationships between system actors, which may be useful for generating important insights for policy design or generating more precise hypotheses for empirical research. Finally, systems modelling approaches are one way of making complex systems analytically tractable by narrowing down on a set of key causal relationships within a system to generate useful predictions and insights about system (Berlow, 2010). Although systems modelling has

been used in the health and education sectors to generate useful predictions, such models have been used more extensively in infrastructure systems research, possibly because the variables are more quantitative in nature and relatively easier to model in comparison to more human or intangible contextual features in health or education.

Table 1: Summary of Macro-systems Approaches with Selected Examples

	Inventory Approaches	**Relational Approaches**	**Systems Modelling**
	Descriptive frameworks that present typologies or lists to define different system components	*Frameworks that specify causal relationships or complementarities between those system components.*	*Frameworks that conceptualize the system through very specific numerical causal relationships.*
Health	· WHO (2007): Characterizes the health system as comprising of 6 functional building blocks (e.g. service delivery, health workforce to name a few)	· Gilson (2003): Characterizes the health system in terms of its stakeholders and trust relationships between them.	· Rwashana et al. (2009): Use dynamic synthesis methodology (DSM) to model the immunization system.
Education	· SABER (2011): Characterizes the education system in terms of thirteen different functions.	· Pritchett (2015): Characterizes the education system in terms of its stakeholders and accountability links between them.	· Kaffenberger and Pritchett (2021): Develop a structural model to capture the dynamics of learning in the education system
Infrastructure	· Rinaldi et al. (2001) outline infrastrucure systems in terms of four main dependencies: physical, cyber, geographical/spatial or logical.	· Saidi et al. (2018): Characterize a multi-layered civil infrastructure system with different interdependencies between physical infrastructure sectors and the broader social economic or political environments.	· Thacker et al. (2017): Develop a network-based *systems-of systems* model for national infrastructure comprising of a group of nodes and edges between system components.

Source: Authors' synthesis

5. Micro-systems approaches

While macro-systems approaches offer big-picture frameworks to understand coherence between many system components and policies, micro-systems approaches focus on the effectiveness of a specific policy – just like impact evaluations. However, the central presumption of micro-systems approaches is that policies cannot be viewed in isolation, but rather need to be designed, implemented, evaluated, and scaled taking the wider context and complementarities within the system into account (Travis et al., 2004; De Savingy & Adam 2009; Snyder, 2013; Pritchett, 2015), and so questions and methods mainly revolve around these issues rather than average treatment effects.

Across the health, education, and infrastructure sectors, a diverse range of analytical approaches self-identify as systems approaches. Each of these approaches are likely to be familiar to readers in some disciplines and unfamiliar to others. They include approaches that aim to help evaluators better understand the roles of mechanisms and contextual factors in producing policy impact, such as realist evaluation (Pawson & Tilley, 1997) and theory-driven evaluation (Coryn et al., 2011), as well as a range qualitative or ethnographic (e.g. George, 2009; Bano & Oberoi, 2020) and mixed method approaches (e.g. Mackenzie et al., 2009; Tuominen et al., 2014) more broadly. They also encompass fields such as implementation science (Rubenstein & Pugh, 2006), some types of meta-analysis and systematic review (e.g. Greenhalgh et al., 2016; Leviton et al., 2017; Masset, 2019), and adaptive approaches to policy design and evaluation (e.g. Andrews et al., 2017). We briefly summarize each of these methods or approaches in this section, before the next section develops a framework to link them back to standard impact evaluation and help prospective systems researchers select among them.

Micro-systems approaches' emphasis on heterogeneity is perhaps best captured by the mantra of the "realist" approach to evaluation, which argues that the purpose of an evaluation should be to identify "what works in which circumstances and for whom?", rather than merely answering the question of "does it work"? (Pawson & Tilley, 1997). More specifically, instead of looking at simple cause and effect relationships, realist research typically aims to develop middle-range theories through developing "context-mechanism-outcome configurations" in which the role of policy context is integral to developing an understanding of how the policy works (Pawson & Tilley, 1997; Greenhalgh et al., 2016). For example, Kwamie et al. (2014) use a realist evaluation to evaluate the impact of the Leadership Development Programme (LDP) delivered to district hospitals in Ghana. Focusing on a district hospital in Accra, they used a range of qualitative data sources to develop causal loop diagrams to explain interactions between contexts, mechanisms, and outcomes. They found that while the training produced some positive short-term outcomes, it was not institutionalised and embedded within the district processes. They argue that this was primarily due to the structure of hierarchical authority in the department, due to which the training was seen as a project coming from the top, and thus reduced initiative on the part of the district managers to institutionalize it.

A related approach is theory-driven evaluation, in which the focus is not just on whether an intervention works but also on its mediating mechanisms – the "why" of impact (Coryn et al., 2011). Theory-driven evaluations take as their starting point the underlying theory of how the policy is intended to achieve its desired outcomes (often expressed in the form of a theory of change diagram), and seek to evaluate each step of this causal process. As with realist approaches, the role of context is critical for theory-driven evaluations, as it is these mechanism-context complementarities that drive heterogeneity of impact across contexts and target populations, and hence the external validity and real-world effectiveness of policies or

interventions. Theory-based and realist evaluations both tend to rely on qualitative methods, either alone or as a supplement to a quantitative impact evaluation (i.e. mixed methods), as limitations of sample size, counterfactual availability, and measurement often make it infeasible to document multiple potential mechanisms quantitatively at the desired levels of nuance and rigor.[6]

Another form of qualitative method widely used by systems researchers is ethnography and participant observation. These are used mainly for the diagnosis of policy problems, refining research hypotheses, or designing new policy interventions, rather than evaluating policy impact *ex post*. For example, George (2009) conducts an ethnographic analysis to examine how formal rules and hierarchies affect informal norms, processes, and power relations in the Indian health system in Koppal state. The study shows that the two key functions of accountability in Koppal's health system – supervision and disciplinary action – are rarely implemented uniformly as these are negotiated by frontline staff in various ways depending on their informal relationships. In the education sector, Bano and Oberoi (2020) use ethnographic methods to understand how innovations are adopted in the context of an Indian NGO that introduced a Teaching at the Right Level (TaRL) intervention, and tease out lessons for how innovations can be scaled and adopted in state systems. In this sense, ethnographic research is a more structured and rigorous version of the informal discussions or anecdotal data that policymakers and evaluators often draw upon in making policy or evaluation decisions, and can be integrated into these processes accordingly (alone or alongside some form of impact evaluation).

[6] Magrath et al. (2019) cite various examples of mixed methods research studies under the Raising Learning Outcomes in Education Systems (RLO) research program.

Systems research often has a specific focus on the implementation, uptake, and scale-up of policy (Hanson, 2015). The discipline of *implementation science* in the health sector, for example, is specifically targeted towards understanding such issues (Rubenstein & Pugh, 2006). Research in implementation science is usually less concerned with the question of what is effective (where there is strong prior evidence on an intervention's efficacy in ideal conditions) and is more concerned with how to implement effectively. Systems researchers who study implementation cater to a set of concerns such as methods for introducing and scaling up new practices, behavior change among practitioners, and the use and effects of patient and implementer participation in improving compliance. Greenhalgh et al. (2017), for example, combine qualitative interviews, ethnographic research, and systematic review to study the implementation of technological innovations in health. They develop the non-adoption, abandonment, scale-up, spread, and sustainability (NASSS) framework to both theorize and evaluate the implementation of health care technologies. Like realist and theory-based evaluation, implementation science research often relies heavily (though not exclusively) on qualitative methods, although these can also be combined with experimental or observational quantitative data.

While these micro-systems approaches are by definition used to analyse the effectiveness of a single policy, some systems researchers have also adapted evidence aggregation methods like systematic reviews and meta-analysis to the interests of systems researchers. While these methods are typically used to summarize impacts or identify an average treatment effect of an intervention by summarizing studies across several contexts, systems researchers focus on using these methods to identify important intervening mechanisms across contexts. For example, Leviton (2017) argues that systematic reviews and meta-analyses can offer bodies of knowledge that support better understanding of external validity by identifying features of

program theory that are consistent across contexts. To identify these systematically, she identifies several techniques to be used in combination with meta-analyses such as a more through description of interventions and their contexts, nuanced theories behind the interventions, and consultation with practitioners. While many of these applications rely on integrating qualitative information into the evidence aggregation process, other researchers use these methods in their traditional quantitative formats but focus specifically on systems-relevant questions of mechanisms, contextual interactions, and heterogeneity. For example, Masset (2019) calculates prediction intervals for various meta-analyses of education interventions and finds that interventions' outcomes are highly heterogeneous and unpredictable across contexts, even for simple interventions like merit-based scholarships. Used in this way, there is methodological overlap between meta-analysis in the systems tradition and how it is commonly used in mainstream impact evaluation. This illustrates one of many ways in which the boundaries between "systems" and "non-systems" research is porous, which both increases the possibilities for productive interchange among research approaches but also creates terminological and conceptual confusion that inhibits it.

Stakeholder mapping or analysis is another method used by systems researchers, to either understand issues of policy implementation or policy design. For example, Sheikh and Porter (2010) conduct a stakeholder analysis to identify key gaps in policy implementation. Using data from in-depth interviews with various stakeholders across five states in India, they highlight bottlenecks in HIV policy implementation (from nine hospitals selected by principles of maximum variation). Like ethnography, stakeholder mapping is an example of a micro-systems approach (because it focuses on the effectiveness of a single policy) but which asks different questions about that policy's effectiveness than standard impact evaluations do.

A final set of micro-systems approaches are grounded in the reality that many questions of policy design and evaluation are situated in complex settings, where policy-context complementarities are so numerous and specific to the contextual setting that the effectiveness of a policy is impossible to predict, for all intents and purposes. Systems researchers argue that for such *complex systems*, which have many "unknown unknowns" with few clear cause and effect relationships, various negative and positive feedback loops and emergent behaviours (Bertalanffy, 1971; Snowden & Boone, 2007), there is a need for a different set of analytical approaches to policy design and evaluation (e.g. Snyder, 2013). This perspective eschews not only the idea of 'best practice' policies but also sometimes the idea of basing adoption decisions on policies' effectiveness in other contexts, because policy dynamics are viewed as so highly context-specific.

A core idea in complex systems theory is that the processes of policy design and implementation should involve an on-going process of iteration with feedback from key stakeholders and decision-makers in the system. For example, Andrews et al. (2013) argue that designing and implementing effective policies for governments in complex settings requires locally driven problem-solving and experimentation, and propose an approach called problem-driven iterative adaptation (PDIA) that emphasizes local problem definition, design, and experimentation. In a different vein, Tsofa et al.'s (2017) "learning sites" approach envisions a long-term research collaboration with a district hospital in which researchers and health practitioners work together over time to uncover and address thorny governance challenges. While the learning site serves to host a series of narrower research studies, the most important elements include formal reflective sessions being regularly held among researchers, between researchers and practitioners, and across learning sites to study complex pathways to change. Such approaches are also closely linked to the *living lab* methodology, which relies on

innovation, experimentation, and participation for diagnosing problems and designing solutions for more effective governance (Dekker et al., 2019).[7]

Table 2: Summary of Micro-systems Approaches

	Realist and Theory-driven evaluations	Ethnographic field studies	Implementation Science	Meta-analysis and systematic reviews	Stakeholder analysis	Approaches for Complex Systems
Health	· Kwamie et al. (2014): A realist evaluation to study the impact of a leadership and management initiative in public health sector in Accra, Ghana.	· George (2009): Ethnographic field studies to study accountability relationships in health systems in India.	· Greenhalgh et al. (2017): Ethnographic research with a systematic review to develop a framework for understanding implementation of health-based technological innovations.	· Kristjansson et al. (2015): Meta-analysis of food supplementation programs on child health showing how place of delivery matters for impact.	· Sheikh and Porter (2009): In-depth stakeholder interviews to understand key gaps in HIV implementation.	· Tsofa et al. (2017): 'Learning sites' approach in which researchers and practitioners work together in a specific geographical space to study decentralization in Kenya.
Education	· Magrath et al. (2019): Quantitative longitudinal data on student outcomes with qualitative interviews to diagnose how accountability functions.	· Bano and Oberoi (2020): Ethnographic study to understand adoption of innovations in an NGO (Pratham) and its lessons for state systems.		· Masset (2019): Meta-analysis of education interventions using prediction intervals (to account for heterogeneity across interventions).		· Crouch and De Stefano (2017): 'Doing Development Differently (DDD)' initiative which emphasizes solutions to be locally designed, owned, implemented, and iterated through repeated cycles.
Infrastructure				· Filazzola et al. (2019): Meta - analysis to study whether green infrastructure is beneficial for biodiversity,	· Tuominen et al. (2014): A method called pluralistic backcasting, in which multiple visions of the future are developed through a participatory and interdisciplinary process.	

Source: Authors' synthesis

6. Systems Approaches and Impact Evaluation

The review of systems approaches in the preceding two sections illustrates the sheer diversity of topics, questions, theories, and methods that can fall within the broad label of systems approaches. It also shows that while systems approaches are sometimes rhetorically positioned in opposition to standard impact evaluation approaches, many of the concerns motivating systems researchers (such as attention to mechanisms, heterogeneity, external validity, implementation and scale-up, the use of qualitative data) can and increasingly are being addressed within the impact evaluation community. At the same time, it is also generally true that systems approaches differ substantially in their prioritization of questions and hence the types of evidence in which they are most interested, so these differences are not purely semantic.

How, then, should a researcher or policymaker think about whether they need to adopt a systems approach to creating and interpreting evidence? And if so, which type of systems approach might be most relevant? In this section, we offer a brief conceptual synthesis and stylized framework to guide thinking on these questions.

For macro-systems approaches, the relationship to standard impact evaluation methods is fairly clear. Macro-systems approaches array the broad range of policies and outcomes relevant to understanding the performance of a given sector, and impact evaluations examine the effect of specific policies on specific outcomes within this framework. Macro-systems frameworks can thus add value to impact evaluation-led approaches to studying policy effectiveness by providing a framework with which to cumulate knowledge, suggesting important variables for impact evaluations to focus on (and potential complementarities among them), and highlighting

gaps in an evidence base. Being more explicit in couching impact evaluations in some kind of broader macro-system framework – whether inventory, relational, or systems modelling – could thus enhance the evidentiary value of systems approaches, as indeed it has begun to do in the systems literatures in the health, education, and infrastructure sectors (e.g. Spivack 2021).

For micro-systems approaches, however, the relationship to (and distinction from) standard impact evaluation methods is more blurry. Among other reasons, this is because our definition of systems approaches as being concerned with multi-dimensional complementarities does not give much guidance as to which types of systems questions and methods might be related to different types of potential complementarities.

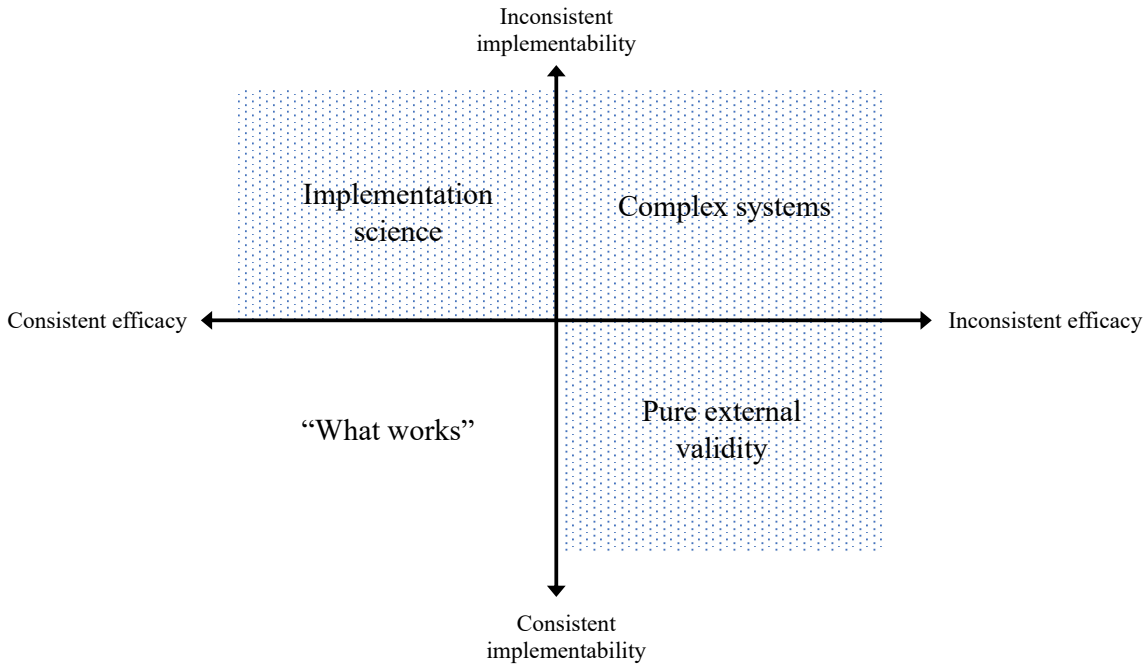
We therefore propose a simple framework that uses a policy's *consistency of implementability* and *consistency of efficacy* to guide choices about the appropriateness of different evidence-creation approaches.[8] By consistency of implementability, we mean the extent to which a given policy can be delivered or implemented correctly (i.e. the desired service delivery outputs can be produced) across a wide range of contexts. Policies whose effective implementation depends on important and numerous complementarities with other policies or aspects of context will tend to have lower consistency of implementability, since these complementary factors will be present in some contexts but not others, whereas policies for whom these complementarities are relatively fewer or less demanding will be able to be implemented more consistently across a wide range of contexts. By consistency of efficacy, we mean the extent to which delivery of a given set of policy outputs results in the same set of outcomes in society across a wide range

[8] Other authors have made similar distinctions among policies with respect to questions of implementation, external validity, and scale-up (e.g. Pritchett and Woolcock, 2004; Bates and Glennerster, 2017; Pritchett, 2017) or with respect to the complexity of problems (Snowden and Boone, 2007). We build on these distinctions and deploy them for a different purpose.

of contexts. As with implementability, policies whose mechanisms rely on many important complementarities with other policies or aspects of context will tend to have lower consistency of efficacy across contexts, and vice versa.

Putting these two dimensions together (Figure 4) yields a set of distinctions among four different stylized types of evidence problems, each of which can be addressed most effectively using different methods for creating and interpreting evidence. In interpreting this diagram, several important caveats are in order. First, this framework is intended to help readers organize the extraordinarily diverse range of micro-systems approaches identified in our review and summarized in our preceding sections, and to identify when they might want to adopt a systems approach and which type might be most useful. But it is not comprehensive taxonomy of all micro-systems approaches, nor do all methods reviewed fit neatly into one category. Second, while we present four stylized "types" of evidence problems for simplicity, the underlying dimensions are continuous spectrums. Finally, complementarities exist and context matters for all policies to at least some extent; the distinctions presented here are intended to be relative in nature, not absolute. With these caveats in mind, we discuss each of these types in turn, highlighting their relationship both to different micro-systems methods as well as to standard impact evaluation approaches.

Figure 4: Synthesizing Micro-systems Approaches



Source: Author's synthesis

The top-left quadrant of Figure 4 corresponds to types of policies which are consistently efficacious across contexts, but which are challenging to implement effectively. We refer to these problems as "implementation science" problems. Handwashing in hospitals is an example of a type of policy that falls in this quadrant, as it is simple and universally effective in reducing hospital-acquired infections but also extremely difficult to get health workers to do routinely. Increasing rates of childhood immunization is another example, as well-established vaccinations are consistently efficacious but many children fail to receive immunizations every year. If a policymaker were considering adopting a policy of promoting vaccinations of children, she ought to be less interested in reading existing evidence (or creating new evidence through research) on the efficacy of the vaccines themselves than in evidence about how to increase vaccination rates.

As discussed in the previous section, implementation science researchers have used a range of methods – qualitative, quantitative, mixed – and theoretical perspectives (e.g. realist evaluation) to address implementation-type problems. Outside of the systems tradition, this concern with the nitty-gritty details of how to better deliver policies and the consequences of minor variations in implementation for take-up is perhaps most closely paralleled by Duflo's (2017) vision of economists (and presumably evidence-creators in other disciplines) as "plumbers" helping governments to improve delivery by varying and evaluating program details. So while implementation is clearly a core focus of many types of systems approaches, this is not to say that researchers who do not self-identify as systems researchers are uninterested in it. That said, systems researchers perhaps tend to be more willing to focus their attention exclusively on implementation issues, as distinct from the policy's impact on final outcomes – a choice which is justifiable for the type of evidence problems posed by policies that share the features of consistent efficacy but inconsistent implementability.

This contrasts to the scenario in the bottom-right quadrant, where a policy is simple to implement but has highly variable efficacy across contexts. This is the classic external validity question: will a policy or intervention that works in one context work in a different context?[9] An example of such a problem is merit-based scholarships for education, which are relatively easy to implement in most contexts, but can have high variance in effectiveness across contexts (Masset, 2019). In terms of methodological responses to such problems, realist and theory-driven evaluations are commonly used by systems researchers to understand these issues of heterogeneous effects and fit with context. Meta-analysis and systematic reviews are also

[9] We call this quadrant "pure" external validity because in practice many impact evaluations (and hence discussions of external validity) combine efficacy and implementation when measuring policy impact or effectiveness, whereas we distinguish between external validity as a matter of a policy's efficacy across contexts (which abstracts from implementation quality) rather than its effectiveness across contexts (which includes implementation quality).

commonly used within the systems tradition to aggregate evidence across studies, but typically with a focus on identifying how context influences policy efficacy more than on estimating an overall average treatment effect, often by supplementing quantitative impact estimates with qualitative data and attention to mechanisms and context (e.g. Greenhalgh et al., 2016; Leviton, 2017). Of course, impact evaluation researchers outside the systems tradition are also increasingly recognizing these issues as important, so once again the difference is largely one of prioritization of questions and of methodological pluralism in addressing them.

Policies which are both inconsistently implementable and inconsistently efficacious fall into the category of *complex systems*. These exhibit features that arise from important and numerous complementarities with other policies and with features of the context, such as: emergent behaviours that are not explained by those interactions in isolation; non-linearities; and system self-organization whilst operating across multiple levels and time periods (Sabelli, 2006). Examples of complex system-type problems in public service delivery include many organization- and sector-level reform efforts, which by their nature affect numerous actors (some of whom are organized and strategic), and depend on the existing state of the system and presence of other related policy interventions. Evidence creation and use takes on very different forms for these type of problems, since knowing that a particular policy worked in another context is unlikely to be informative about its effect in a new context.[10] Evidence generation and learning therefore has to take on very local forms, such as the adaptive experimentation methods (e.g. Andrews et al., 2017) and learning sites and living labs (e.g. Sabel et al., 2012; Tsofa et al., 2017; Dekker et al., 2019) discussed in section 5 above.

[10] The subset of systems studies that view complexity as generating fundamental uncertainty and unpredictability in outcomes (e.g. Sheikh et al., 2011; Snyder, 2013) could be viewed as an extreme case within this quadrant. The underlying epistemological question of whether the outcomes of such systems are impossible to predict or just very difficult to predict is beyond the scope of this article.

Finally, some policies may fall in the bottom-left quadrant of Figure 4 (consistent implementability, consistent efficacy). Such policies are actually relatively amenable to straightforward evaluate-and-transport or evaluate-and-scale-up forms of evidence-based policy, so delving deeply into the complexities of context and broader systems may be unnecessary – or at least not a priority for scarce attention and resources. While context matters for the implementability and efficacy of all policies to some degree, policies such as cash transfers have been shown to be consistently effective in achieving poverty reduction outcomes across a wide range of contexts and are relatively simple to implement. As Bates and Glennerster (2017) note, it is a fallacy to think that all interventions must be re-evaluated in every context in which they are tried, and for policies in this bottom-left quadrant systems approaches might not be necessary at all. Just as there are complex system-type policy problems for which evidence is not generalizable and nearly all learning must be local, there are also "what works"-type policy problems for which evidence is highly generalizable. The challenge for selecting a method of evidence generation and interpretation, then, is being able to predict *ex ante* which type of policy problem one is facing.

How might a researcher or policymaker actually go about deciding which quadrant of this framework they are in when deciding what type of evidence they need in order to make decisions about the adoption and design of a new policy? Several approaches are possible, although each face their own challenges. First, one might approach the question of consistency of implementability and efficacy empirically, by aggregating evidence across multiple contexts and/or target groups through systematic review and meta-analysis. Indeed, multi-intervention meta-analyses such as Vivalt (2020) demonstrate that some interventions exhibit much higher heterogeneity of impact across contexts. Unfortunately, such meta-analyses do not routinely distinguish between implementation and efficacy as causes for this heterogeneity, although in

principle they could – particularly when quantitative methods are supplemented with qualitative data in trying to aggregate evidence about interventions' full causal chains (e.g. Kneale et al., 2018). Second, one could approach the question theoretically, by developing priors about the complexity of each policy's theory of change (i.e. intended mechanism) and its scope for complementarities with other policies or aspects of context in terms of implementation and efficacy. Finally, Williams (2020) proposes a methodology of *mechanism mapping* that combines theory-based and empirics-based approaches to developing predictions about how a policy's mechanism is likely to interact with its context, and thus how heterogeneous its implementability and efficacy are likely to be. All of these approaches have obvious limitations – limited evidence availability, and the difficulty of foreseeing all potential complementarities and their consequences – and in practice would likely need to be combined. Figure 4 is thus likely to be of more use as a conceptual framework or heuristic device than as a device for formally classifying different types of policies. But it may nonetheless help researchers and practitioners structure their thinking about why different types of policies might present different needs in terms of evidence generation.

7. Conclusion

This article has synthesized a wide range of literature that falls under the broad label of systems approaches to public service delivery, drawing key distinctions within it and linking it to standard impact evaluation-led approaches to evidence-based policymaking. Based on our review of studies in health, education, and infrastructure, we have argued that systems approaches are united in their focus on multi-dimensional complementarities between policies and aspects of context as the key challenge for creating and using evidence. This results in a different prioritization of types of questions and greater methodological pluralism, and also

gives rise to a range of different types of systems approaches, each suited to different situations and questions.

Our systems-perspective synthesis in some ways echoes, but goes beyond, discipline-specific attempts to grapple with these issues. In economics, for instance, issues of complementarity among management structures and processes are perhaps the central focus of the field of organizational economics (Brynjolfsson & Milgrom, 2013) as well as common focuses (at least along one or two dimensions) of impact evaluations (Bandiera et al., 2010; Andrabi et al., 2020). Indeed, Besley et al.'s recent (2021) review of the literature on bureaucracy and development (which also calls for a systems perspective) highlights the potential for this literature to draw increasingly organizational economics and industrial organization. Similarly, understanding the impact of policies in general rather than partial equilibrium has long been valued (Acemoglu, 2010) and issues of external validity, implementation, and policy scale-up are now at the forefront of impact evaluation (e.g. Duflo 2017; Vivalt, 2020). In comparative politics, discussion of scope conditions for theories and mixed methods are frequently used to understand mechanisms and heterogeneity (e.g. Falleti et al., 2009). And in public administration, questions around how to incorporate complexity of policy implementation and governance networks in research methods (Klijn, 2008), and new governance approaches to address policy design in the face of such complexity are being increasingly discussed (OECD, 2017).

These convergences of interest, theory, and method present opportunities for cross-sectoral and cross-disciplinary learning. And while these overlaps of questions and methods do serve as a warning against strawman characterizations of other disciplines, so too can they serve to conceal real differences in the specifics of choosing and combining analytical methods, in how

theoretical frameworks are constructed and tested, and – most of all – in the extent to which questions about context and complementarity are prioritized when thinking about policy effectiveness. It is our hope that this article provides readers from a range of backgrounds with a better understanding of the current state of literature on systems approaches, ideas for new avenues of connection with their work, and a common conceptual foundation on which to base dialogue with researchers from different traditions who share the goal of using evidence to improve public service delivery.

References

Acemoglu, Daron. (2010). Theory, General Equilibrium, and Political Economy in Development Economics. *The Journal of Economic Perspectives, 24*(3), 17-32.

Andrabi. Tahir, Das, Jishnu, Khwaja, Asim I, Ozyurt, Selcuk, & Singh, Niharika. (2020). Upping the Ante: The Equilibrium Effects of Unconditional Grants to Private Schools. *The American Economic Review, 110*(10), 3315.

Andrews, Matt, Pritchett, Lant, & Woolcock, Michael. (2013). Escaping Capability Traps Through Problem Driven Iterative Adaptation (PDIA). *World Development, 51*, 234-244.

Andrews, Matt, Pritchett, Lant, & Woolcock, Michael. (2017*). Building state capability: Evidence, analysis, action*. Oxford: Oxford University Press.

Bandiera, Oriana, Barankay, Iwan, & Rasul, Imran. (2010). Social Incentives in the Workplace. *The Review of Economic Studies, 77*(2), 417-458.

Bandiera, Oriana, et al. 2016. State Effectiveness, Growth and Development. IGC, www.theigc.org/wp-content/uploads/2014/09/IGCEvidencePaperState.pdf.

Bano, M., and Oberoi, Z. 2020. Embedding Innovation in State Systems: Lessons from Pratham in India. RISE Working Paper Series. 20/058. https://doi.org/10.35489/BSG-RISE-WP_2020/058.

Bates, M. Ann., and Glennerster, Rachel. 2017. The Generalizability Puzzle. Stanford Social Innovation Review. Retrieved from https://ssir.org/articles/entry/the_generalizability_puzzle.

Bertalanffy, L. (1971). *General system theory : Foundations, development, applications*. London: Allen Lane.

Besley, Timothy, Robin Burgess, Adnan Khan, and Guo Xu. 2021. Bureaucracy and Development. Mimeo, August 5.

Berlow, Eric. (2010). How complexity leads to simplicity, TED Talk. Retrieved from www.ted.com/talks/eric_berlow_how_complexity_leads_to_simplicity.html

Brynjolfsson, Erik, and Paul Milgrom. (2013). Complementarity in Organizations. In Gibbons, Robert, and John Roberts, The Handbook of Organizational Economics, Oxford: Princeton University Press, 11-55.

Bold, Tessa, Kimenyi, Mwangi, Mwabu, Germano, Ng'ang'a, Alice, & Sandefur, Justin. (2018). Experimental evidence on scaling up education reforms in Kenya. *Journal of Public Economics, 168*, 1-20.

Burns, D., & Worsley, S (2015). *Navigating Complexity in International Development: Facilitating sustainable change at scale*. Practical Action Publishing.

Carey, Gemma, Malbon, Eleanor, Carey, Nicole, Joyce, Andrew, Crammond, Brad, & Carey, Alan. (2015). Systems science and systems thinking for public health: A systematic review of the field. *BMJ Open, 5*(12), E009002.

Coryn, Chris L. S, Noakes, Lindsay A, Westine, Carl D, & Schröter, Daniela C. (2011). A Systematic Review of Theory-Driven Evaluation Practice From 1990 to 2009. *American Journal of Evaluation, 32*(2), 199-226.

Deaton, Angus. (2010). Instruments, Randomization, and Learning about Development. *Journal of Economic Literature, 48*(2), 424-455.

Deeg, Richard. (2007). Complementarity and institutional change in capitalist systems. *Journal of European Public Policy, 14*(4), 611-630.

De Savigny, D., & Adam, T. 2009. *Systems thinking for health systems strengthening* (Alliance Flagship report series). Geneva: World Health Organization.

Dekker, Rianne, Franco Contreras, Juan, & Meijer, Albert. (2020). The Living Lab as a Methodology for Public Administration Research: A Systematic Literature Review of its Applications in the Social Sciences. *International Journal of Public Administration, 43*(14), 1207-1217.

Dudenhoeffer, Donald, Permann, May, & Manic, Milos. (2006). CIMS: A framework for infrastructure interdependency modelling and analysis. *Proceedings of the 38th Conference on Winter Simulation, 478-485.*

Duflo, Esther. (2017). The Economist as Plumber. *The American Economic Review, 107*(5), 1-26.

Falleti, Tulia G, & Lynch, Julia F. (2009). Context and Causal Mechanisms in Political Analysis. *Comparative Political Studies, 42*(9), 1143-1166.

Frenk, Julio. (2010). The global health system: Strengthening national health systems as the

next step for global progress. *PLoS Medicine, 7*(1), E1000089.

Filazzola, Alessandro, Shrestha, Namrata, MacIvor, J. Scott, & Stanley, Margaret. (2019). The contribution of constructed green infrastructure to urban biodiversity: A synthesis and meta-analysis. *The Journal of Applied Ecology, 56*(9), 2131-2143.

Gilson, Lucy. (2003). Trust and the development of health care as a social institution. *Social Science & Medicine (1982), 56*(7), 1453-1468.

Gilson, Lucy Ed. 2012. Health Policy and Systems Research - A Methodology Reader. *World Health Organization.* www.who.int/alliance-hpsr/resources/reader/en/.

Greenhalgh, Trisha, Macfarlane, Fraser, Steed, Liz, & Walton, Robert. (2016). What works for whom in pharmacist-led smoking cessation support: Realist review. *BMC Medicine, 14*(1), 209.

Greenhalgh, Trisha, Wherton, Joseph, Papoutsi, Chrysanthi, Lynch, Jennifer, Hughes, Gemma, A'Court, Christine, . . . Shaw, Sara. (2017). Beyond Adoption: A New Framework for Theorizing and Evaluating Nonadoption, Abandonment, and Challenges to the Scale-Up, Spread, and Sustainability of Health and Care Technologies. *Journal of Medical Internet Research, 19(11), E367.*

George, Asha. (2009). 'By papers and pens, you can only do so much': Views about accountability and human resource management from Indian government health administrators and workers. *The International Journal of Health Planning and Management,24*(3), 205-224.

Hall, J., Tran, M., Hickford, A., & Nicholls, R. 2016. *The Future of National Infrastructure: A systems-of-systems approach*. Cambridge University Press.

Hanson, Kara. 2015. What Can Education Systems Research Learn from Health Systems Research? *RISE Working Paper 15/003.* www.gov.uk/dfid-research-outputs/rise-working-paper-15-003-what-can-education-systems-research-learn-from-health-systems-research.

Halsey, Rogers, & Demas, Angela. 2013. The What, Why, and How of the Systems Approach for Better Education Results (SABER). *The World Bank*, 2 Aug. 2013, documents.worldbank.org/curated/en/2013/04/18070354/systems-approach-better-education-results-saber.

Hawe, Penelope. (2015). Lessons from Complex Interventions to Improve Health. *Annual Review of Public Health, 36*(1), 307-323.

Homer, Jack B, & Hirsch, Gary B. (2006). System Dynamics Modeling for Public Health: Background and Opportunities. *American Journal of Public Health (1971), 96*(3), 452-458.

Ichniowski, Casey, & Shaw, Kathryn. (2003). Beyond Incentive Pay: Insiders' Estimates of the Value of Complementary Human Resource Management Practices. *The Journal of Economic Perspectives, 17*(1), 155-180.

Kaffenberger, Michelle, & Pritchett, Lant. (2021). A structured model of the dynamics of student learning in developing countries, with applications to policy. *International Journal of Educational Development, 82*, 102371.

Klijn, Erik-Hans. (2008). Complexity Theory and Public Administration: What's New? *Public Management Review, 10*(3), 299-317.

Kneale, Dylan, Thomas, James, Bangpan, Mukdarut, Waddington, Hugh, & Gough, David. (2018). Conceptualising causal pathways in systematic reviews of international development interventions through adopting a causal chain analysis approach. *Journal of Development Effectiveness, 10*(4), 422-437.

Kristjansson, Elizabeth, Francis, Damian K, Liberato, Selma, Jandu, Maria Benkhalti, Welch, Vivian, Batal, Malek, . . . Petticrew, Mark. (2015). Food Supplementation for Improving the Physical and Psychosocial Health of Socio-economically Disadvantaged Children Aged Three Months to Five Years: A Systematic Review. *Campbell Systematic Review, 11*(1), 1-226

Mackenzie, M, Koshy, P, Leslie, W, Lean, M, & Hankey, C. (2009). Getting beyond outcomes: A realist approach to help understand the impact of a nutritional intervention during smoking cessation. *European Journal of Clinical Nutrition, 63*(9), 1136-1142.

Magrath, Bronwen, Aslam, Monazza, & Johnson, David. (2019). Systems Research in Education: Designs and methods. *Research in Comparative and International Education, 14*(1), 7-29.

Marchal, Bruno, Van Belle, Sara, Van Olmen, Josefien, Hoerée, Tom, & Kegels, Guy. (2012). Is realist evaluation keeping its promise? A review of published empirical studies in the field of health systems research. *Evaluation (London, England. 1995),18*(2), 192-212.

Marchal, Bruno, Westhorp, Gill, Wong, Geoff, Van Belle, Sara, Greenhalgh, Trisha, Kegels, Guy, & Pawson, Ray. (2013). Realist RCTs of complex interventions – An oxymoron. *Social Science & Medicine (1982), 94*, 124-128.

Martineau, Fred P. (2016). People-centred health systems: Building more resilient health systems in the wake of the Ebola crisis. *International Health, 8*(5), 307-309.

Masset, Edoardo. (2019). Impossible generalisations: meta-analyses of education interventions in international development. RISE Programme. Retrieved from https://riseprogramme.org/sites/default/files/inline-files/Masset%2011052019.pdf

Mills, Anne. (2012). Health policy and systems research: Defining the terrain; identifying the methods. *Health Policy and Planning, 27*(1), 1-7.

Moore, Mark. (2015). Creating Efficient, Effective, and Just Educational Systems through Multi-Sector Strategies of Reform. RISE Working Paper 15/004, Blavatnik School of Government, Oxford University, Oxford, U.K. Retrieved from https://riseprogramme.org/sites/default/files/2020-11/RISE_WP-004_Moore-REV%20copy.pdf.

Muller, Sean. (2015). Causal Interaction and External Validity: Obstacles to the Policy Relevance of Randomized Evaluations. *World Bank Economic Review* 29, S217-S225.

Organisation For Economic Co-Operation Development. (2017). *Systems Approaches to Public Sector Challenges: Working with Change*. OECD Publishing.

Ottens, Maarten, Franssen, Maarten, Kroes, Peter, & Van De Poel, Ibo. (2006). Modelling infrastructures as socio-technical systems. *International Journal Of Critical Infrastructures, 2*(2/3), Pp133-145.

Pawson, R., & Tilley, N. 1997. *Realistic evaluation*. London; Thousand Oaks, Calif.: Sage.

Pritchett, Lant, & Woolcock, Michael. (2004). Solutions When the Solution is the Problem: Arraying the Disarray in Development. *World Development, 32*(2), 191-212.

Pritchett, Lant, & Sandefur, Justin. (2015). Learning from Experiments when Context Matters. *The American Economic Review, 105*(5), 471-475.

Pritchett, Lant. (2015). RISE Programme, Working Paper 15/005, Blavatnik School of Government, Oxford University, Oxford, U.K. Retrieved from https://riseprogramme.org/sites/default/files/inline-files/RISE_WP-005_Pritchett_2.pdf.

Pritchett, Lant. (2017). The Evidence' About 'What Works' in Education: Graphs to Illustrate External Validity and Construct Validity. Center for Global Development. Retrieved from https://www.cgdev.org/sites/default/files/evidence-about-what-works-education-graphs-illustrate-external-validity.pdf.

Pritchett, Lant. (2018). What We Learned from Our RISE Baseline Diagnostic Exercise. Rise Programme. Retrieved from www.riseprogramme.org/blog/baseline_diagnostic_exercise_1.

Reiss, Peter C, & Wolak, Frank A. (2007). Chapter 64 Structural Econometric Modeling: Rationales and Examples from Industrial Organization. In *Handbook of Econometrics* (Vol. 6, pp. 4277-4415). Elsevier B.V.

Rinaldi, S.M, Peerenboom, J.P, & Kelly, T.K. (2001). Identifying, understanding, and analyzing critical infrastructure interdependencies. *IEEE Control Systems, 21*(6), 11-25.

Sabelli, Nora H. (2006). Complexity, Technology, Science, and Education. *The Journal of the Learning Sciences, 15*(1), 5-9.

Saidi, Saeid, Kattan, Lina, Jayasinghe, Poornima, Hettiaratchi, Patrick, & Taron, Joshua. (2018). Integrated infrastructure systems—A review. *Sustainable Cities and Society, 36*, 1-11.

Sheikh, Kabir, Gilson, Lucy, Agyepong, Irene Akua, Hanson, Kara, Ssengooba, Freddie, & Bennett, Sara. (2011). Building the field of health policy and systems research: Framing the questions. *PLoS Medicine, 8*(8), E1001073.

Sheikh, Kabir, & Porter, John. (2010). Discursive gaps in the implementation of public health policy guidelines in India: The case of HIV testing. *Social Science & Medicine (1982), 71*(11), 2005-2013.

Snowden, David J, & Boone, Mary E. (2007). A leader's framework for decision making. *Harvard Business Review, 85*(11), 68-149.

Snyder, Sean. (2013). The Simple, the Complicated, and the Complex: Educational Reform Through the Lens of Complexity Theory. *OECD Education Working Papers*, No. 96, OECD Publishing. http://dx.doi.org/10.1787/5k3txnpt1lnr-en

Stacey, R. (2010). *Complexity and organizational reality : Uncertainty and the need to rethink management after the collapse of investment capitalism* (2nd ed.). London: Routledge.

Sturmberg, Joachim P, & Martin, Carmel M. (2009). Complexity and health - yesterday's traditions, tomorrow's future. *Journal of Evaluation in Clinical Practice, 15*(3), 543-548.

Thacker, Scott, Pant, Raghav, & Hall, Jim W. (2017). System-of-systems formulation and disruption analysis for multi-scale critical national infrastructures. *Reliability Engineering & System Safety, 167*, 30-41.

Tsofa, Benjamin, Molyneux, Sassy, Gilson, Lucy, & Goodman, Catherine. (2017). How does decentralisation affect health sector planning and financial management? a case study of early effects of devolution in Kilifi County, Kenya. *International Journal for Equity in Health,16*(1), 151.

Tuominen, Anu, Tapio, Petri, Varho, Vilja, Järvi, Tuuli, & Banister, David. (2014). Pluralistic backcasting: Integrating multiple visions with policy packages for transport climate policy. *Futures: The Journal of Policy, Planning and Futures Studies, 60*, 41-58.

Vivalt, E. 2020. How Much Can We Generalize from Impact Evaluations? *Journal of the European Economic Association* 18(6): 3045–3089.

Williams, Martin. (2020). External validity and policy adaptation: From impact evaluation to policy design. *World Bank Research Observer*, 35(2): 158–191.

World Bank. (2014). SABER in Action: An Overview - Strengthening Education Systems to Achieve Learning for All. Washington, DC: World Bank. Retrieved from https://documents1.worldbank.org/curated/en/866881468323335358/pdf/80059-REVISED-SABER-in-Action-An-Overview.pdf

World Bank. (2018). World Development Report 2018: Learning to Realize Education's Promise. Washington, DC: World Bank. Retrieved from https://www.worldbank.org/en/publication/wdr2018#

World Health Organization. (2004). Strengthening Health Systems: The Role and Promise of Health Policy and Systems Research. Retrieved from www.who.int/alliance-hpsr/resources/publications/hssfr/en/.

World Health Organization. (2007). Everybody's business -- strengthening health systems to improve health outcomes: WHO's framework for action. World Health Organization. Retrieved from https://apps.who.int/iris/handle/10665/43918