5 Will It Work Here? Using Case Studies to Generate 'Key Facts' About Complex Development Programs

Michael Woolcock

> Immersion in the particular proved, as usual, essential for the catching of anything general.
>
> Albert Hirschman[1]

> [T]he bulk of the literature presently recommended for policy decisions . . . cannot be used to identify "what works here". And this is not because it may fail to deliver in some particular cases [; it] is not because its advice fails to deliver what it can be expected to deliver . . . The failing is rather that it is not designed to deliver the bulk of the *key facts* required to conclude that it will work here.
>
> Nancy Cartwright and Jeremy Hardie[2]

5.1 Introduction: In Search of 'Key Facts'

Over the last two decades, social scientists across the disciplines have worked tirelessly to enhance the precision of claims made about the impact of development projects, seeking to formally verify 'what works' as part of a broader campaign for 'evidence-based policy-making' conducted on the basis of 'rigorous evaluations'.[3] In an age of heightened public scrutiny of aid budgets and policy effectiveness, and of rising calls by development

The views expressed in this chapter are those of the author alone, and should not be attributed to the World Bank, its executive directors, or the countries they represent. This chapter refines and updates Woolcock (2013), extending issues with which I have been wrestling for many years. I am grateful for the clarifications, corrections, and insights I have gained over many years from colleagues too numerous to mention, even as, of course, errors of fact and interpretation expressed herein remain solely my own.

[1] Hirschman (1967, p. 3).

[2] Cartwright and Hardie (2012, p. 137); emphasis added.

[3] For present purposes I do not want to engage in philosophical debates about what exactly constitutes a 'fact' (or 'key facts'); such issues are amply discussed in the cases presented in Howlett and Morgan (2010). Here I interpret 'key facts' to mean, pragmatically, "crucially important (but too often overlooked) issues that decision-makers, upon learning that a certain development intervention demonstrably worked 'there', need to take into account when considering whether they too can expect similar results if they adopt this intervention 'here'."

https://doi.org/10.1017/9781108688253.006 Published online by Cambridge University Press

agencies themselves for greater accountability and transparency, it was deemed no longer acceptable to claim success for a project if selected beneficiaries or officials merely expressed satisfaction, if necessary administrative requirements had been upheld, or if large sums had been dispersed without undue controversy. For their part, researchers seeking publication in elite empirical journals, where the primary criteria for acceptance was (and remains) the integrity of one's 'identification strategy' – that is, the methods deployed to verify a causal relationship – faced powerful incentives to actively promote not merely more and better impact evaluations, but methods, such as randomized controlled trials (RCTs) or quasi-experimental designs (QEDs), squarely focused on isolating the singular effects of particular variables. Moreover, by claiming to be adopting (or at least approximating) the 'gold standard' methodological procedures of biomedical science, champions of RCTs in particular could impute to themselves the moral and epistemological high ground as 'the white lab coat guys' of development research.

The heightened focus on RCTs as the privileged basis on which to impute causal claims in development research and project evaluation has been subjected to increasingly trenchant critique,[4] but for present purposes my objective is not to rehearse, summarize, or contribute to these debates per se; it is, rather, to assert that these preoccupations have drained attention from an equally important issue, namely our basis for generalizing any claims about impact from different types of interventions across time, contexts, groups, and scales of operation. If identification and causality are debates about 'internal validity', then generalization and extrapolation are concerns about 'external validity'.[5] It surely matters for the latter that we first have a good handle on the former, but even the cleanest estimation of a given project's impact does not axiomatically provide warrant for confidently inferring that similar results can be expected if that project is scaled up or

[4] See, among others, Cartwright (2007), Deaton (2010), Deaton and Cartwright (2018), Pritchett and Sandefur (2015), Picciotto (2012), Ravallion (2009), and Shaffer (2011). Nobel laureate James Heckman has been making related critiques of "randomization bias" in the evaluation of social policy experiments for more than twenty years. And as Achen (Chapter 3, this volume) stresses, RCTs have a long (and not always glorious) history in policy research, the lessons from which most contemporary advocates of RCTs seem completely unaware of.

[5] The distinctions between construct, internal and external validity form, along with replication, the four core elements of the classic quasi-experimental methodological framework of Cook and Campbell (1979). In later work, Cook (2001) was decidedly more circumspect about the extent to which social scientists (of any kind) can draw empirical generalizations. For those engaged in development policy, Williams (2020) argues that rather than focusing on general external validity concerns, the more specific focus should be on identifying how evidence can be used to more accurately discern whether and how a given intervention might be optimally fitted to a novel context.

https://doi.org/10.1017/9781108688253.006 Published online by Cambridge University Press

replicated elsewhere.[6] Yet too often this is precisely what happens: having expended enormous effort and resources in procuring a clean estimate of a project's impact, and having successfully defended the finding under vigorous questioning at professional seminars and review sessions, the standards for inferring that similar results can be expected elsewhere or when 'scaled up' suddenly drop away markedly. The 'rigorous result', if 'significantly positive', slips all too quickly into implicit or explicit claims that 'we know' the intervention 'works' (even perhaps assuming the status of a veritable 'best practice'), the very 'rigor' of 'the evidence' invoked to promote or defend the project's introduction into a novel (perhaps highly uncertain) context. In short, because an intervention demonstrably worked 'there', we all too often and too confidently presume it will also work 'here'.

Even if concerns about the weak external validity of RCTs/QEDs – or, for that matter, any methodology – of development interventions are acknowledged by most researchers, decision-makers still lack a usable framework by which to engage in the vexing deliberations surrounding whether and when it is at least plausible to infer that a given impact result (positive or negative) 'there' is likely to obtain 'here'. Equally importantly, we lack a coherent system-level imperative requiring decision-makers to take these concerns seriously, not only so that we avoid intractable, nonresolvable debates about the effectiveness of entire portfolios of activity ('community health', 'justice reform') or abstractions ('do women's empowerment programs work?'[7]), but, more positively and constructively, so that we can enter into context-specific discussions about the relative merits of (and priority that should be accorded to) roads, irrigation, cash transfers, immunization, legal reform, etc., with some degree of grounded confidence – that is, on the basis of appropriate metrics, theory, experience, and (as we shall see) trajectories and theories of change.

Though the external validity problem is widespread and vastly consequential for lives, resources, and careers, this chapter's modest goal is not to provide a "tool kit" for "resolving it" but rather to promote a broader conversation about how external validity concerns might be more adequately

[6] The veracity of extrapolating given findings to a broader population in large part turns on sampling quality; the present concern is with enhancing the analytical bases for making comparisons about likely impact between different populations, scales of operation (e.g., pilot projects to national programs), and across time.

[7] The insightful review of 'community driven development' programs by Mansuri and Rao (2012) emphasizes the importance of understanding context when making claims about the effectiveness of such programs (and their generalizability), though it has not always been read this way.

https://doi.org/10.1017/9781108688253.006 Published online by Cambridge University Press

addressed in the practice of development. (Given that the bar, at present, is very low, facilitating any such conversations will be a nontrivial achievement.) As such, this chapter presents ideas to think with. Assessing the extent to which empirical claims about a given project's impact can be generalized is only partly a technical endeavor; it is equally a political, organizational, and philosophical issue, and as such usable and legitimate responses will inherently require extended deliberation in each instance. To this end, the chapter is structured in five sections. Following this introduction, Section 5.2 provides a general summary of selected contributions to the issue of external validity from a range of disciplines and fields. Section 5.3 outlines three domains of inquiry ('causal density', 'implementation capabilities', 'reasoned expectations') that, for present purposes, constitute the key elements of an applied framework for assessing the external validity of development interventions generally, and 'complex' projects in particular. Section 5.4 considers the role analytic case studies can play in responding constructively to these concerns. Section 5.5 concludes.

5.2 External Validity Concerns Across the Disciplines: A Short Tour

Development professionals are far from the only social scientists, or philosophers or scientists of any kind, who are confronting the challenges posed by external validity concerns.[8] Consider first the field of psychology. It is safe to say that many readers of this chapter, in their undergraduate days, participated in various psychology research studies. The general purpose of those studies, of course, was (and continues to be) to test various hypotheses about how and when individuals engage in strategic decision-making, display prejudice toward certain groups, perceive ambiguous stimuli, respond to peer pressure, and the like. But how generalizable are these findings? In a detailed and fascinating paper, Henrich, Heine, and Norenzayan (2010a) reviewed hundreds of such studies, most of which had been conducted on college students in North American and European universities. Despite the limited geographical scope of this sample, most of the studies they reviewed readily inferred (implicitly or explicitly) that their findings were indicative of 'humanity' or reflected something fundamental about 'human nature'. Subjecting these broad claims of generalizability

[8] See, among others, March, Sproull, and Tamuz (1991), Morgan (2012), Ruzzene (2012), and Forrester (2017).

https://doi.org/10.1017/9781108688253.006 Published online by Cambridge University Press

to critical scrutiny (for example, by examining the results from studies where particular 'games' and experiments had been applied to populations elsewhere in the world), Henrich et al. concluded that the participants in the original psychological studies were in fact rather WEIRD – western, educated, industrialized, rich and democratic – since few of the findings of the original studies could be replicated in "non-WEIRD" contexts (see also Henrich, Heine, and Norenzayan 2010b).

Consider next the field of biomedicine, whose methods development researchers are so often invoked to adopt. In the early stages of designing a new pharmaceutical drug, it is common to test prototypes on mice, doing so on the presumption that mouse physiology is sufficiently close to human physiology to enable results for the former to be inferred for the latter. Indeed, over the last several decades a particular mouse – known as 'Black 6' – has been genetically engineered so that biomedical researchers around the world are able to work on mice that are literally genetically identical. This sounds ideal for inferring causal results: biomedical researchers in Norway and New Zealand know they are effectively working on clones, and thus can accurately compare findings. Except that it turns out that in certain key respects mouse physiology is different enough from human physiology to have compromised "years and billions of dollars" (Kolata 2013: A19) of biomedical research on drugs for treating burns, trauma, and sepsis, as reported in a *New York Times* summary of a major (thirty-nine coauthors) paper published in the prestigious *Proceedings of the National Academy of Sciences* (see Seok et al. 2013). In an award-winning science journalism article, Engber (2011) summarized research showing that Black 6 was not even representative of mice – indeed, upon closer inspection, Black 6 turns out to be "a teenaged, alcoholic couch potato with a weakened immune system, and he might be a little hard of hearing." An earlier study published in *The Lancet* (Rothwell 2005) reviewed nearly 200 RCTs in biomedical and clinical research in search of answers to the important question "To whom do the results of this trial apply?" and concluded, rather ominously, that the methodological quality of many of the published studies was such that even their internal validity, let alone their external validity, was questionable. Needless to say, it is more than a little disquieting to learn that even the people who do actually wear white lab coats for a living have their own serious struggles with external validity.[9]

[9] It is worth pointing out that the actual "gold standard" in clinical trials requires not merely the random assignment of subjects to treatment and control groups, but that the allocation be 'triple blind' (i.e.,

https://doi.org/10.1017/9781108688253.006 Published online by Cambridge University Press

Consider next a wonderful simulation paper in health research, which explores the efficacy of two different strategies for identifying the optimal solution to a given clinical problem, a process the authors refer to as "searching the fitness landscape" (Eppstein et al. 2012).[10] Strategy one entails adopting a verified 'best practice' solution: you attempt to solve the problem, in effect, by doing what experts elsewhere have determined is the best approach. Strategy two effectively entails making it up as you go along: you work with others and learn from collective experience to iterate your way to a customized 'best fit'[11] solution in response to the particular circumstances you encounter. The problem these two strategies confront is then itself varied. Initially the problem is quite straight forward, exhibiting what is called a 'smooth fitness landscape' – think of being asked to climb an Egyptian pyramid, with its familiar symmetrical sides. Over time the problem being confronted is made more complex, its fitness landscape becoming increasingly rugged – think of being asked to ascend a steep mountain, with craggy, idiosyncratic features. Which strategy is best for which problem? It turns out the 'best practice' approach is best – but only as long as you are climbing a pyramid (i.e., facing a problem with a smooth fitness landscape). As soon as you tweak the fitness landscape just a little, however, making it even slightly 'rugged', the efficacy of 'best practice' solutions fall away precipitously, and the 'best fit' approach surges to the lead. One can over-interpret these results, of course, but given the powerful imperatives in development to identify "best practices" (as verified, preferably, by an RCT/QED) and replicate "what works," it is worth pondering the implications of the fact that the 'fitness landscapes' we face in development are probably far more likely to be rugged than smooth, and that compelling

neither the subjects themselves, the front-line researchers, nor the principal investigators know who has been assigned to which group until after the study is complete), that control groups receive a placebo treatment (i.e., a treatment that looks and feels like a real treatment, but is in fact not one at all), and that subjects cross over between groups mid-way through the study (i.e., the control group becomes the treatment group, and vice versa) – all to deal with well-understood sources of bias (e.g., Hawthorn effects) that could otherwise compromise the integrity of the study. Needless to say, it is hard to imagine that more than handful of policy intervention, let alone development projects, could come remotely close to upholding these standards.

[10] In a more applied version of this idea, Pritchett, Samji, and Hammer (2012) argue for "crawling the design space" as the strategy of choice for navigating rugged fitness environments.

[11] The concept of 'best fit' comes to development primarily through the work of David Booth (2012); in the Eppstein et al. (2012) formulation, the equivalent concept for determining optimal solutions to novel problems in different contexts emerges through what they refer to as 'quality improvement collaboratives' (QICs). Their study effectively sets up an empirical showdown between RCTs and QICs as rival strategies for complex problem solving.

https://doi.org/10.1017/9781108688253.006 Published online by Cambridge University Press

experimental evidence (supporting a long tradition in the history of science) now suggests that promulgating best practice solutions is a demonstrably inferior strategy for resolving them.

Two final studies demonstrate the crucial importance of implementation and context for understanding external validity concerns in development. Bold et al. (2013) deploy the novel technique of subjecting RCT results themselves to an RCT test of their generalizability using different types of implementing agencies. Earlier studies from India (e.g., Banerjee et al. 2007, Duflo, Dupas, and Kremer 2012, Muralidharan and Sundararaman 2010) famously found that, on the basis of an RCT, contract teachers were demonstrably 'better' (i.e., both more effective and less costly) than regular teachers in terms of helping children to learn. A similar result had been found in Kenya, but as with the India finding, the implementing agent was an NGO. Bold et al. took essentially an identical project design but deployed an evaluation procedure in which 192 schools in Kenya were randomly allocated either to a control group, an NGO-implemented group, or a Ministry of Education-implemented group. The findings were highly diverse: the NGO-implemented group did quite well relative to the control group (as expected), but the Ministry of Education group actually performed *worse* than the control group. In short, the impact of "the project" was a function not only of its design but, crucially and inextricably, of its implementation and context. As the authors aptly conclude, "the effects of this intervention appear highly fragile to the involvement of carefully-selected non-governmental organizations. Ongoing initiatives to produce a fixed, evidence-based menu of effective development interventions will be potentially misleading if interventions are defined at the school, clinic, or village level without reference to their institutional context" (Bold et al. 2013: 7).[12]

A similar conclusion, this time with implications for the basis on which policy interventions might be 'scaled up', emerges from an evaluation of a small business registration program in Brazil (see Bruhn and McKenzie 2013). Intuition and some previous research suggests that a barrier to growth faced by small unregistered firms is that their very informality denies them access to legal protection and financial resources; if ways could be found to lower the barriers to registration – for example, by reducing fees, expanding

[12] See also the important work of Denizer, Kaufmann, and Kraay (2012), who assess the performance of more than 6,000 World Bank projects from inception to completion, a central finding of which is the key role played by high-quality task team leaders (i.e., those responsible for the project's management and implementation on a day-to-day basis) in projects that are not only consistently rated 'satisfactory' but manage to become 'satisfactory' after a mid-term review deeming their project 'unsatisfactory'.

https://doi.org/10.1017/9781108688253.006 Published online by Cambridge University Press

information campaigns promoting the virtues of registration, etc. – many otherwise unregistered firms would surely avail themselves of the opportunity to register, with both the firms themselves and the economy more generally enjoying the fruits. This was the basis on which the state of Minas Gerais in Brazil sought to expand a business start-up simplification program into rural areas: a pilot program that had been reasonably successful in urban areas now sought to 'scale up' into more rural and remote districts, the initial impacts extrapolated by its promoters to the new levels and places of operation. At face value, this was an entirely sensible expectation, one that could also be justified on intrinsic grounds: one could argue that all small firms, irrespective of location, should as a matter of principle be able to register. Deploying an innovative evaluation strategy centered on the use of existing administrative data, Bruhn and McKenzie found that despite faithful implementation the effects of the expanded program on firm registration were net *negative*; isolated villagers, it seems, were so deeply wary of the state that heightened information campaigns on the virtues of small business registration only confirmed their suspicions that the government's real purposes were probably sinister and predatory, and so even those owners that once might have registered their business now did not. If only with the benefit of hindsight, 'what worked' in one place and at one scale of operation was clearly inadequate grounds for inferring what could be expected elsewhere at a much larger one.[13]

In this brief tour[14] of fields ranging from psychology, biomedicine, and clinical health to education, regulation, and criminology we have compelling empirical evidence that inferring external validity to given empirical results – that is, generalizing findings from one group, place, implementation modality, or scale of operation to another – is a highly fraught exercise. As the opening epigraph wisely intones, evidence supporting claims of a significant impact 'there', *even (or especially) when that evidence is a product of a putatively rigorous research design*, does not "deliver the bulk of the key facts required to conclude that it will work here." What might those missing

[13] See also the insightful discussion of the criminology impact evaluation literature in Sampson (2013), who argues strongly for exploring the notion of "contextual causality" as a basis for inferring what might work elsewhere. Lamont (2012) also provides a thoughtful overview of evaluation issues from a sociological perspective.

[14] Econometricians have recently begun to focus more concertedly on external validity concerns (e.g., Allcott and Mullainathan, 2012; Angrist and Fernandez-Val, 2010), though their contributions to date have largely focused on technical problems emergent within evaluations of large social programs in OECD countries (most notably the United States) rather than identifying pragmatic guidelines for replicating or expanding different types of projects in different types of (developing) country contexts.

https://doi.org/10.1017/9781108688253.006 Published online by Cambridge University Press

"key facts" be? Clearly some interventions can be scaled and replicated more readily than others, so how might the content of those "facts" vary between different types of interventions?

In the next section, I propose three categories of issues that can be used to interrogate given development interventions and the basis of the claims made regarding their effectiveness; I argue that these categories can yield potentially useful and usable "key facts" to better inform pragmatic decision-making regarding the likelihood that results obtained 'there' can be expected 'here'. In Section 2.4 I argue that analytic case studies can be a particularly fruitful empirical resource informing the tone and terms of this interrogation, especially for complex development interventions. I posit that this fruitfulness rises in proportion to the 'complexity' of the intervention: the higher the complexity, the more salient (even necessary) inputs from analytic case studies become as contributors to the decision-making process.

5.3 Elements of an Applied Framework for Identifying 'Key Facts'

Heightened sensitivity to external validity concerns does not axiomatically solve the problem of how exactly to make difficult decisions regarding whether, when, and how to replicate and/or scale up (or, for that matter, cancel) interventions on the basis of an initial empirical result, a challenge that becomes incrementally harder as interventions themselves, or constituent elements of them, become more 'complex' (defined below). Even if we have eminently reasonable grounds for accepting a claim about a given project's impact 'there' (with 'that group', at this 'size', implemented by 'these people' using 'this approach'), under what conditions can we confidently infer that the project will generate similar results 'here' (or with 'this group', or if it is 'scaled up', or if implemented by 'those people' deploying 'that approach')? We surely need firmer analytical foundations on which to engage in these deliberations; in short, we need more and better "key facts," and a corresponding theoretical framework able to both generate and accurately interpret those facts.

One could plausibly defend a number of domains in which such "key facts" might reside, but for present purposes I focus on three:[15] 'causal density' (the

[15] These three domains are derived from my reading of the literature, numerous discussions with senior operational colleagues, and my hard-won experience both assessing complex development interventions (e.g., Barron, Diprose, and Woolcock, 2011) and advising others considering their expansion/

https://doi.org/10.1017/9781108688253.006 Published online by Cambridge University Press

extent to which an intervention or its constituent elements are 'complex'); 'implementation capability' (the extent to which a designated organizational entity in the new context can in fact faithfully implement the type of intervention under consideration); and 'reasoned expectations' (the extent to which claims about actual or potential impact are understood within the context of a grounded theory of change specifying what can reasonably be expected to be achieved by when). I address each of these domains in turn.

5.3.1 Causal Density

Conducting even the most routine development intervention is difficult, in the sense that considerable effort needs to be expended at all stages over long periods of time, and that doing so may entail carrying out duties in places that are dangerous ('fragile states') or require navigating morally wrenching situations (dealing with overt corruption, watching children die).[16] If there is no such thing as a 'simple' development project, we need at least a framework for distinguishing between different types and degrees of complexity, since this has a major bearing on the likelihood that a project (indeed, a system or intervention of any kind) will function in predictable ways, which in turn shapes the probability that impact claims associated with it can be generalized.

One entry point into analytical discussions of complexity is of course 'complexity theory', a field to which social scientists engaging with policy issues have increasingly begun to contribute and learn,[17] but for present purposes I will create some basic distinctions using the concept of 'causal density' (see Manzi 2012). An entity with low causal density is one whose constituent elements interact in precisely predictable ways: a wrist watch, for example, may be a marvel of craftsmanship and micro-engineering, but its genius actually lies in its relative 'simplicity': in the finest watches, the cogs comprising the internal mechanism are connected with such a degree of precision that they keep near perfect time over many years, but this is possible because every single aspect of the process is perfectly understood.

replication elsewhere. In the spirit in which this chapter is written, I remain very open to the possibility that other domains should also be considered.

[16] The idea of causal density comes from neuroscience, computing, and physics, and can be succinctly defined as "the number of independent significant interactions among a system's components" (Shanahan, 2008: 041924).

[17] A sampling of this literature across the disciplines includes Byrne (2013), Byrne and Callighan (2013), Colander and Kupers (2014), Ramalingam (2014), and Room (2011).

https://doi.org/10.1017/9781108688253.006 Published online by Cambridge University Press

Development interventions (or aspects of interventions[18]) with low causal density are ideally suited for assessment via techniques such as RCTs because it is reasonable to expect that the impact of a particular element can be isolated and empirically discerned, and the corresponding adjustments or policy decisions made. Indeed, the most celebrated RCTs in the development literature – assessing deworming pills, textbooks, malaria nets, classroom size, cameras in classrooms to reduce teacher absenteeism – have largely been undertaken with interventions (or aspect of interventions) with relatively low causal density. If we are even close to reaching "proof of concept" with interventions such as immunization and iodized salt it is largely because the underlying physiology and biochemistry *has come to be* perfectly understood, and their implementation (while still challenging logistically) requires relatively basic, routinized behavior on the part of front-line agents (see Pritchett and Woolcock 2004). In short, attaining "proof of concept" means the proverbial 'black box' has essentially been eliminated – everything going on inside the 'box' (i.e., the dynamics behind every mechanism connecting inputs and outcomes) is known or knowable.[19]

Entities with high causal density, on the other hand, are characterized by high uncertainty, which is a function of the numerous pathways and feedback loops connecting inputs, actions, and outcomes, the entity's openness to exogenous influences, and the capacity of constituent elements (most notably people) to exercise discretion (i.e., to act independently of or in accordance with rules, expectations, precedent, passions, professional norms, or self-interest). Parenting is perhaps the most familiar example of a high causal density activity. Humans have literally been raising children forever, but as every parent knows, there are often many factors (known and unknown) intervening between their actions and the behavior of their offspring, who are intensely subject to peer pressure and willfully act in accordance with their own (often fluctuating, seemingly quixotic) wishes. Despite millions of years and billions of 'trials', we have not produced anything remotely like "proof of concept" with parenting, even if there are certainly useful rules of thumb. Each generation produces its own bestselling

[18] See Ludwig Kling, and Mullainathan (2011) for a discussion of the virtues of conducting delineated 'mechanism experiments' within otherwise large social policy interventions.

[19] Such knowledge is also readily shareable and cumulative over time. The seminal 2015 paper in physics documenting the existence and weight of the Higgs boson particle, for example, set a "world record" for the number of coauthors: an astounding 5,154 (see Aad et al., 2015). In contrast, books marking the centenary of World War I, perhaps the seminal geopolitical event of the twentieth century, continue to debate lingering points of disagreement, and are mostly written by a single historian.

https://doi.org/10.1017/9781108688253.006 Published online by Cambridge University Press

'manual' based on what it regards as the prevailing scientific and collective wisdom, but even if a given parent dutifully internalizes and enacts the latest manual's every word it is far from certain that his/her child will emerge as a minimally functional and independent young adult; conversely, a parent may know nothing of the book or unwittingly engage in seemingly contrarian practices and yet happily preside over the emergence of a perfectly normal young adult.[20]

Assessing the veracity of development interventions (or aspects of them) with high causal density (e.g., women's empowerment projects, programs to change adolescent sexual behavior in the face of the HIV/AIDS epidemic) requires evaluation strategies tailored to accommodate this reality. Precisely because the 'impact' (wholly or in part) of these interventions often cannot be truly isolated, and is highly contingent on the quality of implementation, any observed impact is very likely to change over time, across contexts, and at different scales of implementation; as such, we need evaluation strategies able to capture these dynamics and provide correspondingly usable recommendations. Crucially, strategies used to assess high causal density interventions are not "less rigorous" than those used to assess their low causal density counterpart; any evaluation strategy, like any tool, is "rigorous" to the extent it deftly and ably responds to the questions being asked of it.[21]

To operationalize causal density we need a basic analytical framework for distinguishing more carefully between these 'low' and 'high' extremes: we can agree that a lawn mower and a family are qualitatively different 'systems', but how can we array the spaces in between?[22] Four questions can be proposed to distinguish between different types of problems in development.[23] First, how many person-to-person transactions are required?[24] Second, how much

[20] Such books are still useful, of course, and diligent parents do well to read them; the point is that at best the books provide general guidance at the margins on particular issues, which is incorporated into the larger storehouse of knowledge the parent has gleaned from their own parents, through experience, common sense, and the advice of significant others.

[21] That is, hammers, saws, and screwdrivers are not "rigorous" tools; they become so to the extent they are correctly deployed in response to the distinctive problem they are designed to solve.

[22] In the complexity theory literature, this space is characteristically arrayed according to whether problems are 'simple', 'complicated', 'complex', and 'chaotic' (see Ramalingam and Jones, 2009). There is much overlap in these distinctions with the framework I present herein, but my concern (and that of the colleagues with whom I work most closely on this) is primarily with articulating pragmatic questions for arraying development interventions, which leads to slightly different categories.

[23] The first two questions (or dimensions) come from Pritchett and Woolcock (2004); the latter two from Andrews, Pritchett, and Woolcock (2017).

[24] Producing a minimally educated child, for example, requires countless interactions between teacher and student (and between students) over many years (roughly 1,000 hours per year of instruction); the

https://doi.org/10.1017/9781108688253.006 Published online by Cambridge University Press

discretion is required of front-line implementing agents?[25] Third, how much pressure do implementing agents face to do something other than respond constructively to the problem?[26] Fourth, to what extent are implementing agents required to deploy solutions from a known menu or to innovate in situ?[27] These questions are most useful when applied to specific operational challenges; rather than asserting that (or trying to determine whether) one 'sector' in development is more or less 'complex' than another (e.g., 'health' versus 'infrastructure'), it is more instructive to begin with a locally nominated and prioritized problem (e.g., how can workers in this factory be afforded adequate working conditions and wages?) and asking of it the four questions posed above to interrogate its component elements. An example of an array of such problems within 'health' is provided in Table 5.1; by providing categorical yes/no answers to these four questions we can arrive at five discrete kinds of problems in development: technocratic, logistical, implementation intensive services, implementation intensive obligations, and complex.

So understood, problems are truly 'complex' that are highly transaction intensive, require considerable discretion by implementing agents, yield powerful pressures for those agents to do something other than implement a solution, and have no known (ex ante) solution.[28] The eventual solutions to these *kinds* of problems are likely to be highly idiosyncratic and context

raising or lowering of interest rates is determined at periodic meetings by a handful of designated technical professionals.

[25] Being an effective social worker requires making wrenching discretionary decisions (e.g., is this family sufficiently dysfunctional that I should withdraw the children and make them wards of the state?); reducing some problems to invariant rules (e.g., the age at which young adults are sufficiently mature to drive, vote, or drink alcohol) should in principle make their implementation relatively straightforward by reducing discretion entirely, but as Gupta (2012) powerfully shows for India, weak administrative infrastructure (e.g., no birth certificates or land registers) can render even the most basic demographic questions (age, number of children, size of land holding) matters for discretionary interpretation by front-line agents, with all the potential for abuse and arbitrariness that goes with it.

[26] Virtually everyone agrees that babies should be immunized, that potholes should be fixed, and that children should be educated; professionals implementing these activities will face little political resistance or 'temptations' to do otherwise (except perhaps just not showing up for work). Those enforcing border patrols, regulating firms, or collecting property taxes, on the other hand, will encounter all manner of resistance and 'temptations' (e.g., bribes, kickbacks) to be less than diligent.

[27] Even when a problem is clear and well understood (e.g., sugary foods, a sedentary lifestyle, and smoking are not good for one's health), it may or may not map onto a known, universal, readily implementable solution.

[28] In more vernacular language we might characterize such problems as 'wicked' (after Churchman, 1967).

https://doi.org/10.1017/9781108688253.006 Published online by Cambridge University Press

Table 5.1 Classification of activities in 'health'

	Local discretion?	Transaction intensive?	Contentious; Tempting alternatives?	Known technology?	Type of implementation challenge
Iodization of salt	No	No	No	Yes	*Technocratic* (policy decree + light implementation)
Vaccinations	No	Yes	No	Yes	*Logistical* (implementation intensive, but 'easy')
Ambulatory curative care	Yes	Yes	No(ish)	Yes	*Implementation Intensive Services* (welcomed, expected)
Regulating private providers	Yes	Yes	Yes	Yes	*Implementation Intensive Obligations* (resisted, evaded)
Promoting preventive health	Yes	Yes	No	No	*Complex* (Implementation intensive, motivation hard, solutions require continuous innovation)

Source: Adapted from Pritchett (2013)

specific; as such, and irrespective of the quality of the evaluation strategy used to discern their 'impact', the default assumption regarding their external validity should be, I argue, zero. Put differently, in such instances the burden of proof should lie with those claiming that the result *is* in fact generalizable. (This burden might be slightly eased for 'implementation intensive' problems, but some considerable burden remains nonetheless.) I hasten to add, however, that this does not mean others facing similarly 'complex' (or

https://doi.org/10.1017/9781108688253.006 Published online by Cambridge University Press

‘implementation intensive’) challenges elsewhere have little to learn from a successful (or failed) intervention’s experiences; on the contrary, it may be highly instructive, but its “lessons” reside less in the content of its final design characteristics than in the processes of exploration and incremental understanding by which a solution was proposed, refined, supported, funded, implemented, refined again, and assessed – that is, in the ideas, principles, and inspiration from which, over time, a solution was crafted and enacted. This is the point at which analytic case studies can demonstrate their true utility, as I discuss in the following sections.

5.3.2 Implementation Capability

Another danger stemming from a single-minded focus on a project’s design characteristics as the causal agent determining observed outcomes is that implementation dynamics are largely overlooked, or at least assumed to be nonproblematic. If, as a result of an RCT (or series of RCTs), a given conditional cash transfer (CCT) program is deemed to have ‘worked’,[29] we all too quickly presume that it can and should be introduced elsewhere, in effect ascribing to it “proof of concept” status. Again, we can be properly convinced of the veracity of a given evaluation’s empirical findings and yet have grave concerns about its generalizability. If from a ‘causal density’ perspective our four questions would likely reveal that in fact any given CCT comprises numerous elements, some of which are ‘complex’, from an ‘implementation capability’ perspective the concern is more prosaic: how confident can we be that any designated implementing agency in the new country or context (e.g., Ministry of Social Welfare) would in fact have the capability to do so, at the designated scale of operation?

Recent research and everyday experience suggests, again, that the burden of proof should lie with those claiming or presuming that the designated implementing agency in the proposed context is indeed up to the task (Pritchett and Sandefur 2015). Consider the delivery of mail. It is hard to think of a less contentious and ‘less complex’ task: everybody wants their mail to be delivered accurately and punctually, and doing so is almost entirely a logistical exercise.[30] The procedures to be followed are unambiguous,

[29] See, among others, the extensive review of the empirical literature on CCTs provided in Fiszbein and Schady (2009); Baird et al. (2013) provide a systematic review of the effect of both conditional and unconditional cash transfer programs on education outcomes.

[30] Indeed, the high-profile advertising slogan of a large, private international parcel service is “We love logistics.”

https://doi.org/10.1017/9781108688253.006 Published online by Cambridge University Press

universally recognized (by international agreement), and entail little discretion on the part of implementing agents (sorters, deliverers). A recent empirical test of the capability of mail delivery systems around the world, however, yielded sobering results. Chong et al. (2014) sent letters to 10 nonexistent addresses in 159 countries, all of which were signatories to an international convention requiring them simply to return such letters to the country of origin (in this case the United States) within 90 days. How many countries were actually able to perform this most routine of tasks? In 25 countries *none* of the 10 letters came back within the designated timeframe; of countries in the bottom half of the world's education distribution the average return rate was 21 percent of the letters. Working with a broader cross-country dataset documenting the current levels and trends in state capability for implementation, Andrews, Pritchett, and Woolcock (2017) ruefully conclude that, by the end of the twenty-first century, only about a dozen of today's low-income countries will have acquired levels of state capability equal to that of today's least-rich OECD countries.[31]

The general point is that in many developing countries, especially the poorest, implementation capability is demonstrably low for 'logistical' tasks, let alone for 'complex' ones. 'Fragile states', almost by definition, cannot readily be assumed to be able to undertake complex tasks (such as responding to medical emergencies after natural disasters) even if such tasks are desperately needed there. And even if they are in fact able to undertake some complex projects (such as regulatory or tax reform), which would be admirable, yet again the burden of proof in these instances should reside with those arguing that such capability to implement the designated intervention does indeed exist (or can readily be acquired). For complex interventions as here defined, high-quality implementation is inherently and inseparably a constituent element of any success they may enjoy (see Honig 2018); the presence in novel contexts of implementing organizations with the requisite capability thus should be demonstrated rather than assumed by those seeking to replicate or expand 'complex' interventions.

5.3.3 Reasoned Expectations

The final domain of consideration, which I call 'reasoned expectations', focuses attention on an intervention's known or imputed trajectory of

[31] An applied strategy for responding to the challenges identified therein is presented in Andrews, Pritchett, and Woolcock (2013, 2017).

https://doi.org/10.1017/9781108688253.006 Published online by Cambridge University Press

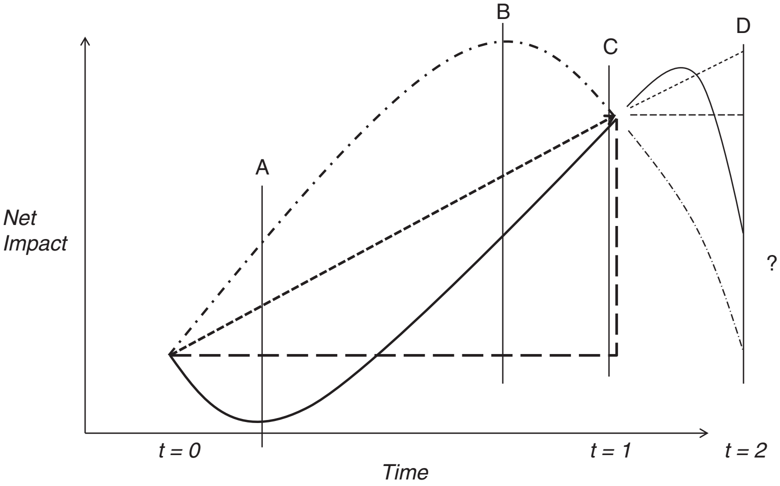

Figure 5.1 Understanding impact trajectories
Source: Woolcock (2013)

change. By this I mean that any empirical claims about a project’s putative impact, *independently of the method(s) by which the claims were determined*, should be understood in the light of where we should reasonably expect a project to be by when. As I have documented elsewhere (Woolcock 2009), the default assumption in the vast majority of impact evaluations is that change over time is monotonically linear: baseline data is collected (perhaps on both a ‘treatment’ and a ‘control’ group) and after a specified time follow-up data is also obtained; following necessary steps to control for the effects of selection and confounding variables, a claim is then made about the net impact of the intervention, and, if presented graphically, is done by connecting a straight line from the baseline scores to the net follow-up scores. The presumption of a straight-line impact trajectory is an enormous one, however, which becomes readily apparent when one alters the shape of the trajectory (to, say, a step-function or a J-curve) and recognizes that the period between the baseline and follow-up data collection is mostly arbitrary (or chosen in accordance with administrative or political imperatives); with variable time frames and nonlinear impact trajectories, however, vastly different accounts can be provided of whether or not a given project is “working.”

Consider Figure 5.1. If one was ignorant of a project impact’s underlying functional form, and the net impact of four projects was evaluated “rigorously” at point C, then remarkably similar stories would be told about these projects’ positive impact, and the conclusion would be that they all

https://doi.org/10.1017/9781108688253.006 Published online by Cambridge University Press

unambiguously "worked." But what if the impact trajectory of these four interventions actually differs markedly, as represented by the four different lines? And what if the evaluation was conducted not at point C but rather at points A or B? At point A one tells four qualitatively different stories about which projects are "working"; indeed, if one had the misfortune to be the team leader on the J-curve project during its evaluation by an RCT at point A, one may well face disciplinary sanction for not merely having "no impact" but for making things worse – as verified by "rigorous evidence"! If one then extrapolates into the future, to point D, it is only the linear trajectory that turns out to yield continued gains; the rest either remain stagnant or decline markedly.

The conclusions reached in an otherwise seminal paper by Casey, Glennerster, and Miguel (2012) embody these concerns. Using an innovative RCT design to assess the efficacy of a 'community driven development' project in Sierra Leone, the authors sought to jointly determine the impact of the project on participants' incomes and the quality of their local institutions. They found "positive short-run effects on local public goods and economic outcomes, but no evidence for sustained impacts on collective action, decision-making, or the involvement of marginalized groups, suggesting that the intervention did not durably reshape local institutions" (2012: 1755). This may well be true empirically, but such a conclusion presumes that incomes and institutions change at the same pace and along the same trajectory; most of what we know from political and social history would suggest that institutional change in fact follows a trajectory (if it has one at all) more like a step-function or a J-curve than a straight line, and that our 'reasoned expectations' against which to assess the effects of an intervention trying to change 'local institutions' should thus be guided accordingly.[32]

Recent work deftly exemplifies the importance of such considerations. Baird, McIntosh, and Özler (2019:182) provide interesting findings from an unconditional cash transfer program in Malawi, in which initially significant declines in teen pregnancy, HIV prevalence, and early marriage turned out, upon a subsequent evaluation conducted two years after the program had concluded, to have dissipated. On the other hand, a conditional cash transfer (CCT) program in the same country offered to girls who were not in school led to "sustained program effects on school attainment, early

[32] On the rising (if belated) awareness among senior researchers of the broader importance of incorporating a theory of change into monitoring and evaluation procedures in development, see Gugerty and Karlan (2018).

https://doi.org/10.1017/9781108688253.006 Published online by Cambridge University Press

marriage, and pregnancy for baseline dropouts receiving CCTs. However, these effects did not translate into reductions in HIV, gains in labor market outcomes, or increased empowerment." Same country, different projects, both with variable nonlinear impact trajectories, and thus different conclusions regarding program effectiveness.[33] One surely needs to have several, sophisticated, contextually grounded theories of change to anticipate and accurately interpret such diverse findings at a given point in time – and especially to inform considerations about the programs' likely effectiveness over time in different country contexts. But, alas, this is rarely the case.[34]

Again, the key point here is not that the *empirical* strategy per se is flawed (it clearly is not – in this instance, in fact, it is exemplary); it is that (a) we rarely have more than two data points on which to base any claims about impact, and, when we do, it can lead to rather different interpretations about impact 'there' (and thus its likely variable impact 'here'); and (b) rigorous (indeed all) results must be interpreted against a theory of change. Perhaps it is entirely within historical experience to see no measurable change on institutions for a decade; perhaps, in fact, one needs to toil in obscurity for two or more decades as the necessary price to pay for any 'change' to be subsequently achieved and discerned;[35] perhaps seeking such change is a highly 'complex' endeavor, and as such has no consistent functional form, or has one that is apparent only with the benefit of hindsight, and is an idiosyncratic product of a series of historically contingent moments and processes (see Woolcock, Szreter, and Rao 2011). In any event, the interpretation and implications of "the evidence" from any evaluation of any intervention is never self-evident; it must be discerned in the light of theory and benchmarked against reasoned expectations, especially when that intervention exhibits high causal density and necessarily requires robust implementation capability.[36]

[33] In the 'complex' development space, see Biddulph (2014) on the impact trajectory of a land titling project in Cambodia, which initially showed spectacular results but over time became so contentious that it led to a breakdown in relations between the World Bank and the Government of Cambodia that lasted several years.

[34] In earlier work, these same authors (Baird, McIntosh, and Ozler, 2011) also showed that different ways of measuring the outcome variables also led to very different interpretations of project impact.

[35] Any student of the history of issues such as civil liberties, gender equality, the rule of law, or human rights surely appreciates this; many changes took centuries to be realized, and many clearly remain unfulfilled.

[36] A horticultural analogy can be invoked to demonstrate this point: no one would claim that sunflowers are "more effective" than acorns if we were to test their "growth performance" over a two-month period. After this time the sunflowers would be six feet high and the acorns would still be dormant underground, with "nothing to show" for their efforts. But we know the expected impact trajectory of sunflowers and oak trees: it is wildly different, and as such we judge (or benchmark) their growth

https://doi.org/10.1017/9781108688253.006 Published online by Cambridge University Press

In the first instance this has important implications for internal validity, but it also matters for external validity, since one dimension of external validity is extrapolation over time. As Figure 5.1 shows, the trajectory of change between the baseline and follow-up points bears not only on the claims made about ‘impact’ but also on the claims made about the likely impact of this intervention in the future. These extrapolations only become more fraught once we add the dimensions of scale and context, as the Braun and McKenzie (2013) and Bold et al. (2013) papers reviewed earlier show. The abiding point for external validity concerns is that decision-makers need a coherent theory of change against which to accurately assess claims about a project’s impact ‘to date’ and its likely impact ‘in the future’; crucially, claims made on the basis of a “rigorous methodology” alone do not solve this problem.

5.3.4 Integrating These Domains into a Single Framework

The three domains considered in this analysis – causal density, implementation capability, and reasoned expectations – comprise a basis for pragmatic and informed deliberations regarding the external validity of development interventions in general and ‘complex’ interventions in particular. While data in various forms and from various sources can be vital inputs into these deliberations (see Bamberger, Rao, and Woolcock 2010; Woolcock 2019), when the three domains are considered as part of a single integrated framework for engaging with ‘complex’ interventions, it is extended deliberations on the basis of analytic case studies, I argue, that have a particular comparative advantage for eliciting the “key facts” necessary for making hard decisions about the generalizability of those interventions (or their constituent elements). Indeed, it is within the domains of causal density, implementation capability, and reasoned expectations, I argue, that the “key facts” themselves reside.

These deliberations move from the analytical and abstract to the decidedly concrete when hard decisions have to be made about the impact and generalizability of claims pertaining to truly complex development interventions, such as those seeking to empower the marginalized, enhance the legitimacy of justice systems, or promote more effective local government. The Sustainable Development Goals have put issues such as these squarely and

performance over time accordingly. Unfortunately, we have no such theory of change informing most assessments of most development projects at particular points in time; in the absence of such theories – whether grounded in evidence and/or experience – of multiple data points, and of corresponding trajectories of change, we assume linearity (which for ‘complex’ interventions as defined in this chapter is almost assuredly inaccurate).

https://doi.org/10.1017/9781108688253.006 Published online by Cambridge University Press

Table 5.2 An integrated framework for assessing external validity claims

	Iodization of salt	Vaccinations	Ambulatory curative care	Regulating private providers	Promoting preventive health
Local discretion?	No	No	Yes	Yes	Yes
Transaction intensive?	No	Yes	Yes	Yes	Yes
Contentious; Tempting alternatives?	No	No	No	Yes	No
Known technology?	Yes	Yes	Yes	Yes	Yes
Type of implementation challenge	*Technocratic* Policy decree + light implementation	*Logistical* Implementation intensive, but 'easy'	*Implementation Intensive Services* Welcomed, expected	*Implementation Intensive Obligations* Resisted, evaded	*Complex* Implementation intensive, motivation hard, solutions require continuous innovation
Likelihood impact claims can be scaled, replicated	High →				Low
Utility of case studies in external validity deliberations	Low →				High

Source: Revised from Woolcock (2013)

formally on the global agenda, and in the years leading up to 2030 there will surely be a flurry of brave attempts to 'measure' and 'demonstrate' that all countries have indeed made 'progress' on them. Is fifteen years (2015–2030) a 'reasonable' timeframe over which to expect any such change to occur? What 'proven' instruments and policy strategies can domestic and international actors wield in response to such challenges? There aren't any, and there never will be, at least not in the way there are now 'proven' ways in which to build durable roads in high rainfall environments, tame high inflation, or immunize babies against polio. But we do have an array of tools in the social science kit that can help us navigate the distinctive challenges posed by truly complex problems – we just need to forge and protect the political space in which they can be ably deployed. Analytic case studies, so understood, are one of those tools.

5.4 Harnessing the Distinctive Contribution of Analytic Case Studies

When carefully compiled and conveyed, case studies can be instructive for policy deliberations across the analytic space set out in Table 5.2. Our focus here is on development problems that are highly complex, require robust implementation capability, and unfold along nonlinear context-specific trajectories, but this is only where the comparative advantage of case studies is strongest (and where, by extension, the comparative advantage of RCTs for engaging with external validity issues is weakest). It is obviously beyond the scope of this chapter to provide a comprehensive summary of the theory and strategies underpinning case study analysis,[37] but three key points bear some discussion (which I provide below): the distinctiveness of case studies as a method of analysis in social science beyond the familiar qualitative/quantitative divide; the capacity of case studies to elicit causal claims and generate testable hypotheses; and (related) the focus of case studies on exploring and explaining mechanisms (i.e., identifying how, for whom, and under what conditions outcomes are observed – or "getting inside the black box").

The rising quality of the analytic foundations of case study research has been one of the underappreciated (at least in mainstream social science) methodological advances of the last few decades (Mahoney 2007). Where everyday discourse in development research typically presumes a rigid and binary 'qualitative' or 'quantitative' divide, this is a distinction many contemporary social scientists (especially historians, historical sociologists, and comparative political scientists) feel does not aptly accommodate their work – if 'qualitative' is primarily understood to mean ethnography, participant observation, and interviews. These researchers see themselves as occupying a distinctive epistemological space, using case studies (across varying units of analysis: countries to firms to events) to interrogate instances of phenomena – with an 'N' of, say, 30, such as revolutions – that are "too large" for orthodox qualitative approaches and "too small" for orthodox quantitative analysis. (There is no inherent reason, they argue, why the problems of the world should array themselves in accordance with

[37] Such accounts are provided in the key works of Ragin and Becker (1992), Stake (1995), Burawoy (1998), George and Bennett (2005), Levy (2008), and Yin (2017); see also the earlier work of Ragin (1987) on 'qualitative comparative analysis' and Bates et al. (1998) on 'analytic narratives' (updated in Levy and Weingast, Chapter 11, this volume), and the most recent methodological innovations outlined in Goertz and Mahoney (2012), Gerring (2017), and Goertz (2017).

https://doi.org/10.1017/9781108688253.006 Published online by Cambridge University Press

the bimodal methodological distribution social scientists otherwise impose on them.)

More ambitiously, perhaps, case study researchers also claim to be able to draw causal inferences (see Mahoney 2000; Levy 2008; Cartwright, Chapter 2 this volume). Defending this claim in detail requires engagement with philosophical issues beyond the scope of this chapter,[38] but a pragmatic application can be seen in the law (Honoré 2010), where it is the task of investigators to assemble various forms and sources of evidence (inherently of highly variable quality) as part of the process of building a "case" for or against a charge, which must then pass the scrutiny of a judge or jury: whether a threshold of causality is reached in this instance has very real (in the real world) consequences. Good case study research in effect engages in its own internal dialogue with the 'prosecution' and 'defense', posing alternative hypotheses to account for observed outcomes and seeking to test their veracity on the basis of the best available evidence. As in civil law, a "preponderance of the evidence" standard[39] is used to determine whether a causal relationship has been established. This is the basis on which causal claims (and, needless to say, highly 'complex' causal claims) affecting the fates of individuals, firms, and governments are determined in courts every day; deploying a variant on it is what good case study research entails.

Finally, by exploring 'cases within cases' (thereby raising or lowering the instances of phenomena they are exploring), and by overtly tracing the evolution of given cases over time within the context(s) in which they occur, case study researchers seek to document and explain the processes by which, and the conditions under which, certain outcomes are obtained. (This technique is sometimes referred to as process tracing – or, as noted earlier, assessing the 'causes of effects' as opposed to the 'effects of causes' approach characteristic of most econometric research.) Case study research finds its most prominent place in applied development research and program assessment in the literature on 'realist evaluation',[40] where the abiding focus is exploiting, exploring, and explaining variance (or standard deviations): that is, on identifying what works for whom, when, where, and why.[41] In

[38] But see the discussion in Cartwright and Hardie (2012); Freedman (2008) and especially Goertz and Mahoney (2012) are also instructive on this point. On the significance of "one or a few cases" for advancing theory, see Rueschemeyer (2003) and Small (2009).

[39] In criminal law, of course, the standard is higher: the evidence must be "beyond a reasonable doubt."

[40] The foundational text is Pawson and Tilly (1997).

[41] This strand of work can reasonably be understood as a qualitative complement to Ravallion's (2001) clarion call for development researchers to "look beyond averages."

https://doi.org/10.1017/9781108688253.006 Published online by Cambridge University Press

their study of service delivery systems across the Middle East and North Africa, Brixi, Lust, and Woolcock (2015) use this strategy – deploying existing household survey data to 'map' broad national trends in health and education outcomes, complementing it with analytical case studies of specific locations that are positive 'outliers' – to explain how, within otherwise similar (and deeply challenging) policy environments, some implementation systems become and remain so much more effective than others (see also McDonnell 2020). This is the signature role that case studies can play for understanding, and sharing the lessons from, 'complex' development interventions on their own terms, as has been the central plea of this chapter.

5.5 Conclusion

The energy and exactitude with which development researchers debate the veracity of claims about 'causality' and 'impact' (internal validity) has yet to inspire corresponding firepower in the domain of concerns about whether and how to 'replicate' and 'scale up' interventions (external validity). Indeed, as manifest in everyday policy debates in contemporary development, the gulf between these modes of analysis is wide, palpable, and consequential: the fates of billions of dollars, millions of lives, and thousands of careers turn on how external validity concerns are addressed, and yet too often the basis for these deliberations is decidedly shallow.

It does not have to be this way. The social sciences, broadly defined, contain within them an array of theories and methods for addressing both internal and external validity concerns; they are there to be deployed if invited to the table (see Stern et al. 2012). This chapter has sought to show that 'complex' development interventions require evaluation strategies tailored to accommodate that reality; such interventions are square pegs which when forced into methodological round holes yield confused, even erroneous, verdicts regarding their effectiveness 'there' and likely effectiveness 'here'. In the early twenty-first century, development professionals routinely engage with issues of increasing 'complexity': consolidating democratic transitions, reforming legal systems, promoting social inclusion, enhancing public sector management[42] – the list is endless. These types of

[42] So et al. (2018) use case studies to explain the array of outcomes associated with efforts to reform the public sector in eight East Asian countries. Such massive, contentious, long-term efforts to modernize administrative systems that enable federal governments to function on a day-to-day basis are quintessentially 'complex': one simply cannot conclude that a singular approach did or did not "work;" it is

https://doi.org/10.1017/9781108688253.006 Published online by Cambridge University Press

issues are decidedly (wickedly) 'complex', and responses to them need to be prioritized, designed, implemented, and assessed accordingly. Beyond evaluating such interventions on their own terms, however, it is as important to be able to advise front-line staff, senior management, and colleagues working elsewhere about when and how the "lessons" from these diverse experiences can be applied. Deliberations centered on causal density, implementation capability, and reasoned expectations have the potential to usefully elicit, inform, and consolidate this process.

References

Aad, G., Abbott, B., Abdallah, J. et al. (2015) "Combined measurement of the Higgs boson mass in pp collisions at sqrt[s]= 7 and 8 TeV with the ATLAS and CMS experiments," *Physical Review Letters*, 114(19), 191803. (This paper has 5,154 coauthors, so for obvious reasons not all are listed here.)

Allcott, H. and Mullainathan, S. (2012) External validity and partner selection bias. Cambridge, MA: National Bureau of Economic Research Working Paper No. 18373.

Andrews, M., Pritchett, L., and Woolcock, M. (2013) "Escaping capability traps through problem-driven iterative adaption (PDIA)," *World Development*, 51(11), 234–244.

Andrews, M., Pritchett, L., and Woolcock, M. (2017) *Building state capability: Evidence, analysis, action* New York: Oxford University Press.

Angrist, J. and Fernandez-Val, I. (2010) Extrapolate-ing: External validity and overidentification in the LATE framework. Cambridge, MA: National Bureau of Economic Research, NBER Working Paper 16566.

Baird, S., Ferreira, F., Özler, B., and Woolcock, M. (2013) "Relative effectiveness of conditional and unconditional cash transfers for schooling outcomes in developing countries: A systematic review," *Campbell Systematic Reviews*, 9(8), 1–124.

Baird, S., McIntosh, C., and Özler, B. (2011) "Cash or condition? Evidence from a cash transfer experiment," *Quarterly Journal of Economics*, 126(4), 1709–1753.

Baird, S., McIntosh, C., and Özler, B. (2019) "When the money runs out: Do cash transfers have sustained effects on human capital accumulation?" *Journal of Development Economics*, 140(September), 169–185.

Bamberger, M., Rao, V., and Woolcock, M. (2010) "Using mixed methods in monitoring and evaluation: Experiences from international development" in Tashakkori, A. and Teddlie, C.

unreasonable to expect such a verdict, and one certainly can't lament that verdicts are "merely anecdotal" because an "RCT" of such reforms wasn't conducted (on such reform it would be neither possible nor desirable to take this approach, even if tiny slices of it perhaps could be so interrogated). Rather, all sorts of contingent events and processes aligned to drive observed outcomes in each case; those contemplating public sector reforms in their own country, we argue, are best served (and prepared) by learning from detailed, analytically informed accounts of the diverse experiences of others with 'similar enough' country/political/administrative characteristics.

https://doi.org/10.1017/9781108688253.006 Published online by Cambridge University Press

(eds.) *Handbook of mixed methods in social and behavioral research* (2nd revised edition). Thousand Oaks, CA: Sage Publications, pp. 613–641.

Banerjee, A. V., Cole, S., Duflo, E., and Linden, L. (2007) "Remedying education: Evidence from two randomized experiments in India," *Quarterly Journal of Economics*, 122(3), 1235–1264.

Barron, P., Diprose, R., and Woolcock, M. (2011) *Contesting development: Participatory projects and local conflict dynamics in Indonesia*. New Haven, CT: Yale University Press.

Bates, R., Greif, A., Levi, M., Rosenthal, J. L., and Weingast, B. R. (eds.) (1998) *Analytic narratives*. Princeton, NJ: Princeton University Press.

Biddulph, R. (2014) Cambodia's land management and administration project. Helsinki: WIDER, Working Paper No. 2014/086.

Bold, T., Kimenyi, M., Mwabu, G., Ng'ang'a, A., and Sandefur, J. (2013) Scaling-up what works: Experimental evidence on external validity in Kenyan education. Washington: Center for Global Development, Working Paper No. 321.

Booth, D. (2012) "Aid effectiveness: Bring country ownership (and politics) back in," *Conflict, Security and Development*, 12(5), 537–558.

Brixi, H., Lust, E., and Woolcock, M. (2015) *Trust, voice and incentives: Learning from local success stories in service delivery in the Middle East and North Africa*. Washington, DC: The World Bank.

Bruhn, M. and McKenzie, D. (2013) "Using administrative data to evaluate municipal reforms: An evaluation of the impact of Minas Fácil Expresso," *Journal of Development Effectiveness*, 5(3), 319–338.

Burawoy, M. (1998) "The extended case method," *Sociological Theory*, 16(1), 4–33.

Byrne, D. (2013) "Evaluating complex social interventions in a complex world," *Evaluation*, 19(3), 217–228.

Byrne, D. and Callighan, G. (2013) *Complexity theory and the social sciences: The state of the art*. London: Routledge.

Cartwright, N. (2007) "Are RCTs the gold standard?" *Biosocieties*, 2(2), 11–20.

Cartwright, N. and Hardie, J. (2012) *Evidence-based policy: A practical guide to doing it better*. New York: Oxford University Press.

Casey, K., Glennerster, R., and Miguel, E. (2012) "Reshaping institutions: Evidence on aid impacts using a pre-analysis plan," *Quarterly Journal of Economics*, 127(4), 1755–1812.

Chong, A., La Porta, R., Lopez-de-Silanes, F., and Shleifer, A. (2014) "Letter grading government efficiency," *Journal of the European Economic Association*, 12(2), 277–298.

Churchman, C. W. (1967) "Wicked problems," *Management Science*, 14(4), 141–142.

Colander, D. and Kupers, R. (2014) *Complexity and the art of social science: Solving society's problems from the bottom up*. Princeton, NJ: Princeton University Press.

Cook, T. D. (2001) "Generalization: Conceptions in the social sciences" in Smelser, N. J., Wright, J., and Baltes, P. B. (eds.) *International encyclopedia of the social and behavioral sciences*. Amsterdam: Elsevier, vol. 9, pp. 6037–6043.

Cook, T. D. and Campbell, D. T. (1979) *Quasi-experimentation: Design and analysis issues for field settings*. Boston, MA: Houghton Mifflin Company.

Deaton, A. (2010) "Instruments, randomization, and learning about development," *Journal of Economic Perspectives*, 48(June), 424–455.

Deaton, A. and Cartwright, N. (2018) "Understanding and misunderstanding randomized controlled trials," *Social Science & Medicine*, 210(2018), 2–21.

https://doi.org/10.1017/9781108688253.006 Published online by Cambridge University Press

Denizer, C., Kaufmann, D., and Kraay, A. (2012) "*Good projects or good countries? Macro and micro correlates of World Bank project performance*," Washington, DC: The World Bank, Policy Research Working Papers No. 5646.

Duflo, E., Dupas, P., and Kremer, M. (2012) School governance, teacher incentives, and pupil-teacher ratios: Experimental evidence from Kenyan primary schools. NBER Working Paper No. 17939.

Engber, D. (2011) "The mouse trap (part I): The dangers of using one lab animal to study every disease," *Slate*, November 15. Available at: www.slate.com/articles/health_and_science/the_mouse_trap/2011/11/the_mouse_trap.html (accessed January 22, 2020).

Eppstein, M. J., Horbar, J. D., Buzas, J. S., and Kauffman, S. (2012) "Searching the clinical fitness landscape," *PLoS One*, 7(11), e49901.

Fiszbein, A. and Schady, N. (2009) *Conditional cash transfers: Reducing present and future poverty*. Washington, DC: The World Bank.

Forrester, J. (2017) *Thinking in cases*. Cambridge: Polity Press.

Freedman, D. A. (2008) "On types of scientific enquiry: The role of qualitative reasoning" in Box-Steffensmeier, J., Brady, H. E., and Collier, D. (eds.) *The Oxford handbook of political methodology*. New York: Oxford University Press, pp. 300–318.

George, A. and Bennett, A. (2005) *Case studies and theory development in the social sciences*. Cambridge, MA: MIT Press.

Gerring, J. (2017) *Case study research: Principles and practices* (2nd ed.) New York: Cambridge University Press.

Goertz, G. (2017) *Multimethod research, causal mechanisms, and case studies*. Princeton, NJ: Princeton University Press.

Goertz, G. and Mahoney, J. (2012) *A tale of two cultures: Qualitative and quantitative research in the social sciences*. Princeton, NJ: Princeton University Press.

Gugerty, M. K. and Karlan, D. (2018) *The Goldilocks challenge: Right-fit evidence for the social sector*. New York: Oxford University Press.

Gupta, A. (2012) *Red tape: Bureaucracy, structural violence and poverty in India*. Durham and London: Duke University Press.

Henrich, J., Heine, S. J., and Norenzayan, A. (2010a) "The weirdest people in the world?" *Behavioral and Brain Sciences*, 33(2–3), 61–83.

Henrich, J., Heine, S. J., and Norenzayan, A. (2010b) "Beyond WEIRD: Towards a broad-based behavioral science," *Behavioral and Brain Sciences*, 33(2–3), 111–135.

Hirschman, A. O. (1967) *Development projects observed*. Washington, DC: Brookings Institution.

Honig, D. (2018) *Navigation by judgment: Why and when top-down management of foreign aid doesn't work*. New York: Oxford University Press.

Honoré, A. (2010) "Causation in the law," *Stanford Encyclopedia of Philosophy*. Available at http://stanford.library.usyd.edu.au/entries/causation-law/ (accessed March 20, 2013).

Howlett, P. and Morgan, M. (eds.) (2010) *How well do facts travel? The dissemination of reliable knowledge*. New York: Cambridge University Press.

Kolata, G. (2013) "Mice fall short as test subjects for humans' deadly ills," *The New York Times*, February 11, A19.

Lamont, M. (2012) "Toward a comparative sociology of valuation and evaluation," *Annual Review of Sociology*, 38(1), 201–221.

https://doi.org/10.1017/9781108688253.006 Published online by Cambridge University Press

Levy, J. S. (2008) "Case studies: Types, designs and logics of inference," *Conflict Management and Peace Studies*, 25(1), 1–18.

Ludwig, J., Kling, J. R., and Mullainathan, S. (2011) "Mechanism experiments and policy evaluations," *Journal of Economic Perspectives*, 25(3), 17–38.

Mahoney, J. (2000) "Strategies of causal inference in small-N analysis," *Sociological Methods & Research*, 28(4), 387–424.

Mahoney, J. (2007) "Qualitative methodology and comparative politics," *Comparative Political Studies*, 40(2), 122–144.

Mansuri, G. and Rao, V. (2012) *Localizing development: Does participation work?* Washington, DC: The World Bank.

Manzi, J. (2012) *Uncontrolled: The surprising payoff of trial and error for business, politics, and society*. New York: Basic Books.

March, J. G., Sproull, L. S., and Tamuz, M. (1991) "Learning from samples of one or fewer," *Organization Science*, 2(1), 1–13.

McDonnell, E. M. (2020) *Patchwork Leviathan: Pockets of bureaucratic effectiveness in developing states*. Princeton, NJ: Princeton University Press.

Morgan, M. (2012) "Case studies: One observation or many? Justification or discovery?" *Philosophy of Science*, 79(5), 667–677.

Muralidharan, K. and Sundararaman, V. (2010) *Contract teachers: Experimental evidence from India*. San Diego, CA: Mimeo.

Pawson, R. and Tilly, N. (1997) *Realist evaluation*. London: Sage Publications.

Picciotto, R. (2012) "Experimentalism and development evaluation: Will the bubble burst?" *Evaluation*, 18(2), 213–229.

Pritchett, L. (2013) *The folk and the formula: Fact and fiction in development*. Helsinki: WIDER Annual Lecture 16.

Pritchett, L., Samji, S., and Hammer, J. (2012) It's all about MeE: Using structured experiential learning ('e') to crawl the design space. Helsinki: UNU-WIDER Working Paper No. 2012/104.

Pritchett, L. and Sandefur, J. (2015) "Learning from experiments when context matters," *American Economic Review*, 105(5), 471–475.

Pritchett, L. and Woolcock, M. (2004) "Solutions when the solution is the problem: Arraying the disarray in development," *World Development*, 32(2), 191–212.

Pritchett, L., Woolcock, M., and Andrews, M. (2013) "Looking like a state: Techniques of persistent failure in state capability for implementation," *Journal of Development Studies*, 49(1), 1–18.

Ragin, C. C. (1987) *The comparative method: Moving beyond qualitative and quantitative strategies*. Berkeley and Los Angeles: University of California Press.

Ragin, C. C. and Becker, H. (eds.) (1992) *What is a case? Exploring the foundations of social inquiry*. New York: Cambridge University Press.

Ramalingam, B. (2014) *Aid on the edge of chaos: Rethinking international cooperation in a complex world*. New York: Oxford University Press.

Ramalingam, B. and Jones, H. (with Toussaint Reba and John Young) (2009) Exploring the science of complexity: Ideas and implications for development and humanitarian efforts. London: ODI Working Paper No. 285.

Ravallion, M. (2001) "Growth, inequality and poverty: Looking beyond averages," *World Development*, 29(11), 1803–1815.

https://doi.org/10.1017/9781108688253.006 Published online by Cambridge University Press

Ravallion, M. (2009) "Should the randomistas rule?" *Economists' Voice*, 6(2), 1–5.

Room, G. (2011) *Complexity, institutions and public policy: Agile decision-making in a turbulent world*. Northampton, MA: Edward Elgar Publishing.

Rothwell, P. M. (2005) "External validity of randomized controlled trials: 'To whom do the results of this trial apply?'" *The Lancet*, 365(9453), 82–93.

Rueschemeyer, D. (2003) "Can one or a few cases yield theoretical gains?" in Mahoney, J. and Rueschemeyer, D. (eds.) *Comparative historical analysis in the social sciences*. New York: Cambridge University Press, pp. 305–336.

Ruzzene, A. (2012) "Drawing lessons from case studies by enhancing comparability," *Philosophy of the Social Sciences*, 42(1), 99–120.

Sampson, R. (2013) "The place of context: A theory and strategy for criminology's hard problems," *Criminology*, 51(1), 1–31.

Seok, J., Warren, H. S., Cuenca, A. G., et al. (2013) "Genomic responses in mouse models poorly mimic human inflammatory diseases," *Proceedings of the National Academy of Sciences*, 110(9), 3507–3512. (This paper has 39 coauthors, so for obvious reasons not all of them are listed here.)

Shaffer, P. (2011) "Against excessive rhetoric in impact assessment: Overstating the case for randomised controlled experiments," *Journal of Development Studies*, 47(11), 1619–1635.

Shanahan, M. (2008) "Dynamical complexity in small-world networks of spiking neurons," *Physical Review E*, 78(4), 041924.

Small, M. L. (2009) "How many cases do I need? On science and the logic of case selection in field-based research," *Ethnography*, 10(1), 5–38.

So, S., Woolcock, M., April, L., Hughes, C., and Smithers, N. (eds.) (2018) *Alternative paths to public financial management and public sector reform: Experiences from East Asia*. Washington, DC: The World Bank.

Stake, R. E. (1995) *The art of case study research*. Thousand Oaks, CA: Sage Publications.

Stern, E., Stame, N., Mayne, J., Forss, K., Davies, R., and Befani, B. (2012) Broadening the range of designs and methods for impact evaluation. London: DFID Working Paper No. 38.

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

Woolcock, M. (2009) "Toward a plurality of methods in project evaluation: A contextualized approach to understanding impact trajectories and efficacy," *Journal of Development Effectiveness*, 1(1), 1–14.

Woolcock, M. (2013) "Using case studies to explore the external validity of complex development interventions," *Evaluation*, 19(3), 229–248.

Woolcock, M. (2019) "Reasons for using mixed methods in the evaluation of complex projects" in Nagatsu, M. and Ruzzene, A. (eds.) *Philosophy and interdisciplinary social science: A dialogue*. London: Bloomsbury Academic, pp. 149–171.

Woolcock, M., Szreter, S., and Rao, V. (2011) "How and why does history matter for development policy?" *Journal of Development Studies*, 47(1), 70–96.

Yin, R. K. (2017) *Case study research: Design and methods* (6th ed.) Thousand Oaks, CA: Sage Publications. (First edition published in 1984.)

https://doi.org/10.1017/9781108688253.006 Published online by Cambridge University Press