Using case studies to explore the external validity of 'complex' development interventions

Michael Woolcock

Evaluation 2013 19: 229
DOI: 10.1177/1356389013495210

Article

Evaluation
19(3) 229–248

DOI: 10.1177/1356389013495210
evi.sagepub.com
SAGE

Using case studies to explore the external validity of 'complex' development interventions

Michael Woolcock
World Bank, USA

Abstract
Rising standards for accurately inferring the impact of development projects has not been matched by equivalently rigorous procedures for guiding decisions about whether and how similar results might be expected elsewhere. These 'external validity' concerns are especially pressing for 'complex' development interventions, in which the explicit purpose is often to adapt projects to local contextual realities and where high quality implementation is paramount to success. A basic analytical framework is provided for assessing the external validity of complex development interventions. It argues for deploying case studies to better identify the conditions under which diverse outcomes are observed, focusing in particular on the salience of contextual idiosyncrasies, implementation capabilities and trajectories of change. Upholding the canonical methodological principle that questions should guide methods, not vice versa, is required if a truly rigorous basis for generalizing claims about likely impact across time, groups, contexts and scales of operation is to be discerned for different kinds of development interventions.

Keywords
case studies, complexity, development, evaluation, external validity

> [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. (Cartwright and Hardie, 2012: 137)

Introduction

Over the last 15 years or so, researchers 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

Corresponding author:
Michael Woolcock, Mailstop MC3-306, The World Bank, 1818 H Street NW, Washington, DC 20433, USA.
Email: mwoolcock@worldbank.org

a broader campaign for 'evidence-based policy making' conducted on the basis of 'rigorous evaluations'. Though most development projects for most of the last 50 years have, upon completion, been subjected to some form of review, by the late 1990s the standards typically deployed in doing so were increasingly deemed inadequate: in an age of heightened public scrutiny of aid budgets and policy effectiveness, and of rising calls by development agencies themselves for greater accountability and transparency, it was no longer acceptable to claim 'success' for a project if selected beneficiaries or officials expressed satisfaction, if necessary administrative requirements had been upheld, or if large sums had been dispersed without undue controversy. For their part, researchers seeking publications in elite empirical journals, where the primary criteria for acceptance was (and remains) the integrity of one's 'identification strategy' – i.e. the methods deployed to verify a causal relationship – faced powerful incentives to actively promote not merely more and better impact evaluations, but methods squarely focused on isolating the singular effects of particular variables, such as randomized control trials (RCTs) or quasi-experimental designs (QEDs). Moreover, by claiming to be adopting (or at least approximating) the 'gold standard' methodological procedures of biomedical science, champions of RCTs in particular imputed 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 in recent years,[1] but for present purposes my objective is not to rehearse, summarize or contribute to these debates per se but rather to assert that these preoccupations have drained attention from an equally important issue, namely our basis for generalizing any claims about impact 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'.[2] 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 replicated elsewhere.[3] 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', translates all too quickly into implicit or explicit claims that the intervention now has 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, wherein it is confidently assumed that it will also now 'work'.

These tendencies are reflected in and reinforced by the logic of claim-making surrounding 'systematic reviews' (e.g. the Cochrane and Campbell Collaborations), in which only a tiny fraction of the studies conducted on a particular intervention (i.e. those conducted using an RCT or perhaps a QED) are deemed sufficiently rigorous for determining the 'true' impact of a class of interventions.[4] The very rationale of systematic reviews is to ensure that 'busy policymakers' tasked with making difficult choices under hard time and budget constraints have access to 'warehouses' of verified 'instruments' from which they can prudently chose. In development policy deliberations, especially those premised on identifying interventions most likely to meet predetermined targets (such as the Millennium Development Goals), asking whether and how expectations and project design characteristics might need to be modified for qualitatively different times, places and circumstances is at best a third order consideration; everyone might claim to agree that 'context matters' and that 'one size doesn't fit all', but the prestige and power in most development agencies, large and small, remain squarely with project designers, funders and those granting the project's

initial approval. In recent years this august group has been joined by those given (or assuming) the mantle of determining whether that project – or, more ambitiously, the broader class of interventions ('microfinance', 'agricultural extension') of which the project is a member – actually 'works'.

Even if concerns about the weak external validity of RCTs/QEDs – or for that matter any methodology – are acknowledged by most researchers, development professionals still lack a useable 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, non-resolvable debates about the effectiveness of entire portfolios of activity ('community health', 'justice reform') or abstractions ('do women's empowerment programs work?'[5]) 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 – i.e. on the basis of appropriate metrics, theory, experience and (as we shall see) trajectories of change.

Though the external validity problem is widespread and vastly consequential for lives, resources and careers, my modest goal in this article 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 addressed in the practice of development. (Given that the bar, at present, is very low, facilitating any such conversations will be a non-trivial achievement.) As such, this is an article to think with. Assessing the extent to which empirical claims about a given project's impact can be generalized is only partly a technical endeavour; it is equally a political, organizational and philosophical issue, and as such useable and legitimate responses will inherently require extended deliberation in each instance. To this end, the article is structured in five sections. Following this introduction, section two provides a general summary of selected contributions to the issue of external validity from a range of disciplines. Section three 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 four considers the role analytic case studies can play in responding constructively to these concerns. Section five concludes.

External validity concerns across the disciplines: A short tour

Development professionals are far from the only social scientists, or scientists of any kind, who are confronting the challenges posed by external validity concerns. Consider first the field of psychology. It is safe to say that many readers of this article, 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 towards 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 et al. (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 to critical scrutiny (e.g. 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 et al., 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 mice physiology is sufficiently close to human physiology to enable results on 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 mice 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 (39 co-authors) paper published recently in the prestigious *Proceedings of the National Academy of Sciences* (see Seok et al., 2013). In an award-winning science journalism article, Engber (2012) 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.[6]

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).[7] 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'[8] 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 overinterpret these results, of course, but given the powerful imperatives in development to identify 'best practices' (as verified 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 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 et al., 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 the 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, 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' (p. 7).[9]

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 programme 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 – e.g. by reducing fees, expanding 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 programme into rural areas: a pilot programme 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 programme 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 purpose was 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.[10]

In this brief tour[11] 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 – i.e. 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 'key facts'

be? 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 useable 'key facts' to better inform pragmatic decision-making regarding the likelihood that results obtained 'there' can be expected 'here'. In section 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; indeed, I will argue that this fruitfulness rises in proportion to the 'complexity' of the intervention: in short, the higher the complexity the more salient (even necessary) analytic case studies become.

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' (see 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 'those guys' using 'that 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 'these guys' deploying 'this 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:[12] 'causal density' (the extent to which an intervention or its constituent elements are 'complex'); 'implementation capability' (the extent to which a designated organization 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.

'Causal density'[13]

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). 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 have increasingly begun to contribute and learn (see Byrne, this volume; Byrne and Callaghan; 2013), 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 very genius is its relative 'simplicity': in the finest watches, the cogs comprising the internal mechanism are connected with a degree of

precision such that they keep near perfect time over many years, but this is possible because every single aspect of the process is perfectly understood – the watchmakers have achieved what philosophers call 'proof of concept'. Development interventions (or aspects of interventions[14]) 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 discerned, and the corresponding adjustments or policy decisions made. Indeed, the most celebrated RCTs in the development literature – assessing de-worming 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 only basic, routinized behaviour – see baby, insert needle – on the part of front-line agents (see Pritchett and Woolcock, 2004). In short, when we have 'proof of concept' we have essentially eliminated the proverbial 'black box' – everything going on inside the 'box' (i.e. every mechanism connecting inputs and outcomes) is known or knowable.

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 behaviour of their offspring, who are intensely subject to peer pressure and willfully act in accordance with their own (often fluctuating) 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 best-selling '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.[15]

Assessing the veracity of development interventions (or aspects of them) with high causal density – e.g. women's empowerment projects, programmes to change adolescent sexual behaviour 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 useable 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.[16]

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?[17] Four questions

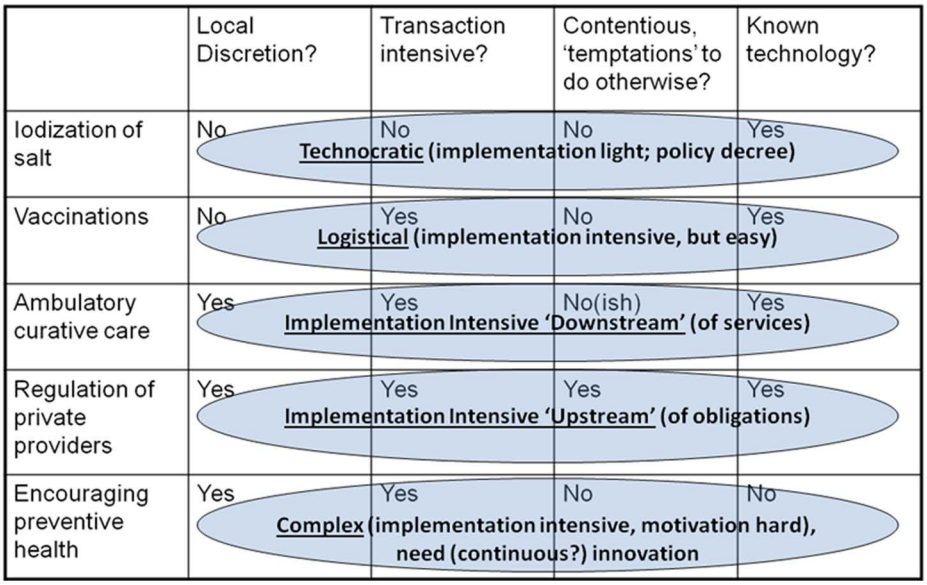

Figure 1. Classification of activities in 'health'.
Adapted from Pritchett (2013).

can be proposed to distinguish between different types of problems in development.[18] First, how many person-to-person transactions are required?[19] Second, how much discretion is required of front-line implementing agents?[20] Third, how much pressure do implementing agents face to do something other than respond constructively to the problem?[21] Fourth, to what extent are implementing agents required to deploy solutions from a known menu or to innovate in situ?[22] 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 Figure 1; by providing straightforward yes/no answers to these four questions we can arrive at five coherent kinds of problems in development: technocratic, logistical, implementation intensive 'downstream', implementation intensive 'upstream', 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.[23] Solutions to these *kinds* of problems are likely to be highly idiosyncratic and context specific; as such, and irrespective of the quality of the evaluation strategy used to discern their 'impact', the default assumption regarding their external validity, I argue, should be 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 'implementation intensive') challenges elsewhere have little to learn from a successful (or failed)

intervention's experiences; on the contrary, it can be highly instructive, but its 'lessons' reside less in the quality of its final 'design' characteristics than the processes of exploration and incremental understanding by which a solution was proposed, refined, supported, funded, implemented, refined again, and assessed – i.e. in the ideas, principles and inspiration from which a solution was crafted and enacted. This is the point at which analytic case studies can demonstrate their true utility, as I discuss below.

'Implementation capability'

Another danger stemming from a single-minded focus on a project's 'design' as the causal agent determining observed outcomes is that implementation dynamics are largely overlooked, or at least assumed to be non-problematic. If, as a result of an RCT (or series of RCTs), a given conditional cash transfer (CCT) programme is deemed to have 'worked',[24] 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 would in fact have the capability to do so?

Recent research (Pritchett et al., 2013) 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. 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 on time, and doing so is almost entirely a logistical exercise[25] – the procedures to be followed are unambiguous, 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. (2012) sent letters to ten deliberately non-existent 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 USA) 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 dataset, Pritchett (2013) calculates that these countries will take roughly 160 years to have post offices with the capability of countries such as Finland and Colombia (which returned 90% of the letters).[26]

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' such as Haiti, almost by definition, cannot readily be assumed to be able to undertake complex tasks (such as disaster relief) even if such tasks are most 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 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; 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.

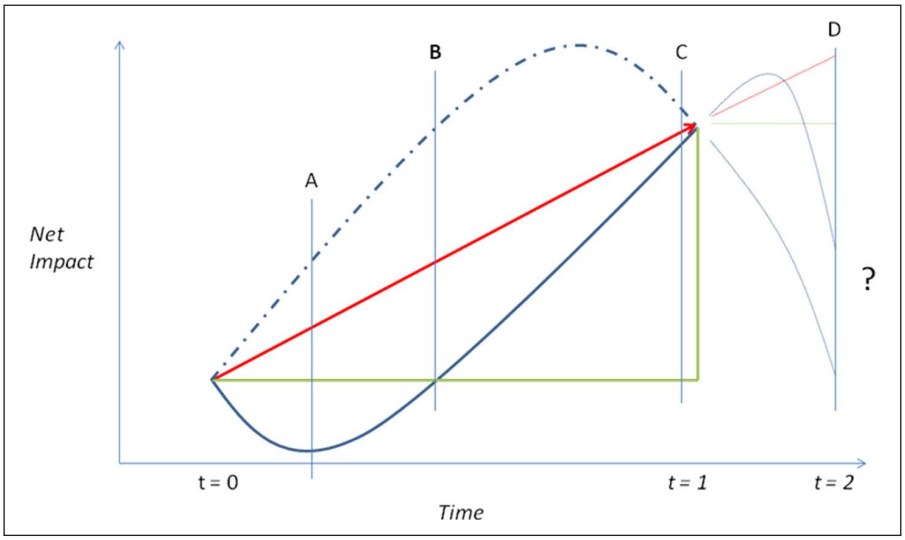

Figure 2. Understanding impact trajectories.

'Reasoned expectations'

The final domain of consideration, which I call 'reasoned expectations', focuses attention on an intervention's known or imputed trajectory of 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 factor out 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 become 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; with variable time frames and non-linear impact trajectories, vastly different accounts can be provided of whether a given project is 'working' or not.

Consider Figure 2. 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 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 working 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.

A recent paper by Casey et al. (2012) embodies 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.' 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 (see Woolcock et al., 2011), and that our 'reasoned expectations' against which to assess the effects of an intervention trying to change 'local institutions' should thus be guided accordingly. Perhaps it is entirely within historical experience to see no measureable 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;[27] perhaps seeking such change is a highly 'complex' endeavour, 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). 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.[28]

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 2 shows, the trajectory of change between the baseline and follow-up points bears not only on the claims made about 'impact' but 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.

Integrating these domains into a single framework. The three domains considered in this analysis – causal density, implementation capability, 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 et al., 2010), 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 delivering the 'key facts' necessary for making hard decisions about the generalizability of those interventions (or their constituent elements).

Considered together (see Figure 3), it should now be apparent that generalizing about projects that exhibit high causal density, require high implementation capability and generate impacts along an unknown (perhaps even unknowable, ex ante) trajectory is a decidedly high-uncertainty undertaking. These three domains are often interrelated – highly complex projects, by their nature, are

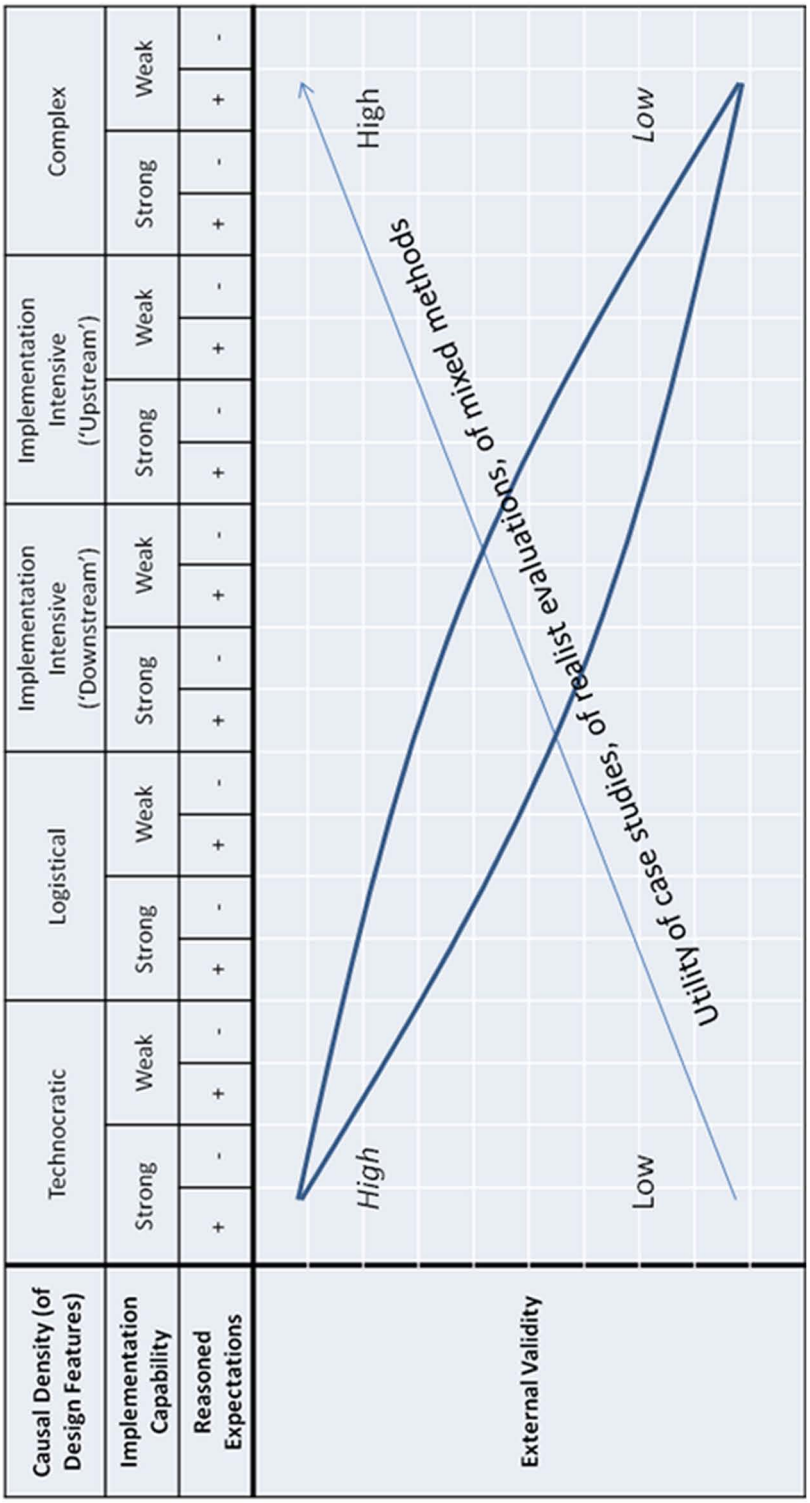

Figure 3. An integrated framework for assessing external validity.

likely to exhibit different impact trajectories in different contexts and/or when implemented by different agencies – but for decision-making purposes they can be considered discrete realms of deliberation. As the next section shows, carefully assembled analytic case studies – in conjunction with mixed method research designs (Bamberger et al., 2010) and realist evaluation strategies (Pawson, 2006) – can be an informed basis on which these deliberations are conducted.

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 Figure 3. Our focus here is on development problems that are highly complex, require robust implementation capability and that unfold along non-linear 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 is weakest). It is obviously beyond the scope of this article to provide a comprehensive summary of the theory and strategies underpinning case study analysis,[29] 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 twenty years (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 accommodates their work, if 'qualitative' is 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 the bi-modal 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). Defending this claim in detail requires engagement with philosophical issues beyond the scope of this article,[30] 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[31] 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, and 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 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 development research and programme assessment in the literature on 'realist evaluation' (the foundational text is Pawson and Tilly, 1997), where the abiding focus is exploiting, exploring and explaining variance (or standard deviations): i.e. on identifying what works for whom, when, where and why.[32] This is the signature role that case studies can play for understanding 'complex' development interventions in particular on their own terms, as has been the central plea of this article.

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 fate 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 article 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'. History is now demanding that development professionals engage with issues of increasing 'complexity': consolidating democratic transitions, reforming legal systems, promoting social inclusion, enhancing public sector management. These types of 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 elicit, inform and consolidate this process.

Acknowledgements

The views expressed in this article are those of the author alone, and should not be attributed to the World Bank, its executive directors or the countries they represent. This article is part of a larger project on identifying practical strategies for assessing the conditions under which the impacts of complex development interventions can be assessed and generalized. I am grateful to Elliot Stern for inviting me to turn some initial thoughts into a more substantial article, to Arathi Rao for diligent research assistance, to April Harding, Heather Lanthorn, Massoud Moussavi, and Lant Pritchett for constructive written comments on an early draft, and to seminar participants at the University of Copenhagen, the World Bank, InterAction, USAID, DFID, SIDA (Sweden), Unite for Sight, the American Sociological Association (Economic Development section

meeting) and the American Evaluation Association for helpful feedback, questions and suggestions. Scott Guggenheim, April Harding, Christopher Nelson, Lant Pritchett, Vijayendra Rao and Sanjeev Sridharan have been valued interlocutors on these issues for many years (though of course the usual disclaimers apply).

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Notes

1. See, among others, Cartwright (2007), Deaton (2010), 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 over 20 years.
2. 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 more recent work, Cook (2001) is decidedly more circumspect about the extent to which social scientists (of any kind) can draw empirical generalizations.
3. 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 programmes) and across time.
4. In a recent systematic review to which I contributed, our team assessed the effectiveness of conditional and unconditional cash transfer programmes, applying (by rule) the RCT-only criteria. This meant that fully 97% of the published literature – more than 4000 studies, many of them published in leading peer-reviewed journals by seasoned practitioners and researchers – had to be declared inadmissible, as essentially having nothing of substance to contribute. Note that this is not a criticism of systematic reviews (or RCTs/QEDs) per se – they are what they are; rather, my concern is the broader apparatus of institutional decision-making that has created, in effect, a monopoly on what counts as a question and what counts as an answer in the assessment of social interventions (with, I would argue, all the attendant inefficiencies one characteristically associates with monopolies).
5. The insightful and instructive review of 'community driven development' programmes by Mansuri and Rao (2012) emphasizes the importance of understanding context when making claims about the effectiveness of such programmes (and their generalizability), though it has not always been read this way.
6. 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. neither the subjects themselves, the front-line researchers nor the principal investigators knows 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 the treatment group becomes the control group) – 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 any policy intervention, let alone a development project, could come remotely close to upholding these standards.
7. In a more applied version of this idea, Pritchett et al. (2012) argue for 'crawling the design space' as the strategy of choice for navigating rugged fitness environments.
8. The concept of 'best fit' comes to development primarily through the work of David Booth (2011); 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.

9. See also the important work of Denizer et al. (2012), who assess the performance of more than 6000 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'.
10. 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.
11. Rao and Woolcock (forthcoming) provide a more extensive review of the literature on external validity and its significance for development policy. 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 programmes in OECD countries (most notably the USA) rather than identifying pragmatic guidelines for replicating or expanding different types of projects in different types of (developing) country contexts.
12. 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 et al., 2011) and advising others considering their expansion/replication elsewhere.
13. 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). More formally, and within economics, it is an extension of the notion of 'Granger causality', in which data from one time-series is used to make predictions about another.
14. See Klinger et al. (2011) for a discussion of the virtues of conducting delineated 'mechanism experiments' within otherwise large social policy interventions.
15. 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.
16. 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.
17. 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 below, 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.
18. The first two questions (or dimensions) come from Pritchett and Woolcock (2004); the latter two from Andrews et al. (forthcoming).
19. Producing a minimally educated child, for example, requires countless interactions between teacher and student (and between students) over many years; the raising or lowering of interest rates is determined at periodic meetings by a handful of designated technical professionals.
20. 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 frontline agents, with all the potential for abuse and arbitrariness that goes with it.
21. 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. 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) to be less than diligent.

22. Even when a problem is clear and well understood – e.g. fatty 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.
23. In more vernacular language we might characterize such problems as 'wicked' (after Churchman, 1967).
24. See, among others, the extensive review of the empirical literature on CCTs provided in Fiszbein and Schady (2009); Baird et al. (2013) provide a more recent 'systematic review' of the effect of both conditional and unconditional cash transfer programmes on education outcomes.
25. Indeed, the high-profile advertising slogan of a large, private international parcel service is: 'We love logistics'.
26. For a broader conceptual and empirical discussion of the evolving organizational capabilities of developing countries see Pritchett et al. (2013). An applied strategy for responding to the challenges identified therein is presented in Andrews et al. (forthcoming).
27. Any student of the history of issues such as civil liberties, gender equality, the rule of law and human rights surely appreciates this; many changes took centuries to be realized, and many remain unfulfilled.
28. In a blog post I have used a horticultural analogy 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 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 – and corresponding trajectories of change, we assume linearity (which for 'complex' interventions as defined in this article is almost assuredly inaccurate).
29. Such accounts are provided in the canonical works of Ragin and Becker (1992), George and Bennett (2005), Gerring (2007) and Yin (2009); see also the earlier work of Ragin (1987) on 'qualitative comparative analysis' and Bates et al. (1998) on 'analytic narratives', and the most recent methodological innovations outlined in Goertz and Mahoney (2012).
30. But see the discussion in Cartwright and Hardie (2012); Freedman (2008) and especially Goertz and Mahoney (2012) are also instructive on this point.
31. In criminal law the standard is higher; the evidence must be 'beyond a reasonable doubt'.
32. 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'.

References

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 (forthcoming) Escaping capability traps through problem-driven iterative adaption (PDIA). *World Development*.

Angrist J and Fernandez-Val I (2010) Extrapolate-ing: external validity and overidentification in the LATE framework. Cambridge, MA: National Bureau of Economic Research. *National Bureau of Economic Research* Working Paper No. 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. London: 3ie.

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 (eds), *Handbook of Mixed Methods in Social and Behavioral Research*, 2nd revised edition. Thousand Oaks, CA: SAGE, 613−41.

Banerjee AV, Cole S, Duflo E and Linden L (2007) Remedying education: evidence from two randomized experiments in India. *Quarterly Journal of Economics* 122(3): 1235−64.

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 BR (1998) *Analytic Narratives*. Princeton, NJ: Princeton University Press.

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 (2011) Aid effectiveness: bring country ownership (and politics) back in. London: ODI Working Paper No. 336.

Bruhn M and McKenzie D (2013) Using administrative data to evaluate municipal reforms: an evaluation of the impact of Minas Fácil Expresso. Washington: World Bank Policy Research Working Paper No. 6368.

Byrne D (2013) Evaluating complex social interventions in a complex world. *Evaluation* 19(3).

Byrne D and Callaghan 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.

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): 1755−812.

Cartwright N and Hardie J (2012) *Evidence-Based Policy: A Practical Guide to Doing it Better*. New York: Oxford University Press.

Chong A, La Porta R, Lopez-de-Silanes F and Shleifer A (2012) Letter grading government efficiency. Cambridge, MA: NBER Working Paper No. 18268.

Churchman CW (1967) Wicked problems. *Management Science* 14(4): 141−2.

Cook TD (2001) Generalization: conceptions in the social sciences. In: Smelser NJ, Wright J and Baltes PB (eds), *International Encyclopedia of the Social and Behavioral Sciences*, Vol. 9. Amsterdam: Elsevier, 6037−43.

Cook TD and Campbell DT (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): 424−55.

Denizer C, Kaufmann D and Kraay A (2012) Good projects or good countries? Macro and micro correlates of World Bank project performance. Washington: 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. URL: http://www.slate.com/articles/health_and_science/the_mouse_trap/2011/11/the_mouse_trap.html

Eppstein MJ, Horbar JD, Buzas JS 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: World Bank.

Freedman DA (2008) On types of scientific enquiry: the role of qualitative reasoning. In: Box-Steffensmeier J, Brady HE and Collier D (eds), *The Oxford Handbook of Political Methodology*. New York: Oxford University Press, 300–18.

George A and Bennett A (2005) *Case Studies and Theory Development in the Social Sciences*. Cambridge, MA: MIT Press.

Gerring J (2007) *Case Study Research: Principles and Practices*. New York: Cambridge 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.

Gupta A (2012) *Red Tape: Bureaucracy, Structural Violence and Poverty in India*. Durham, MD and London: Duke University Press.

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

Henrich J, Heine SJ and Norenzayan A (2010b) Beyond WEIRD: towards a broad-based behavioral science. *Behavioral and Brain Sciences* 33(2−3): 111−35.

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

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

Lamont M (2012) Toward a comparative sociology of valuation and evaluation. *Annual Review of Sociology* 38: 201−21.

Ludwig J, Kling JR 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): 122−44.

Mansuri G and Rao V (2012) *Localizing Development: Does Participation Work?* Washington, DC: World Bank.

Manzi J (2012) *Uncontrolled: The Surprising Payoff of Trial and Error for Business, Politics, and Society*. New York: Basic Books.

Muralidharan K and Sundararaman V (2010) Contract teachers: experimental evidence from India. University of California, San Diego: Mimeo. URL: http://www.fas.nus.edu.sg/ecs/events/seminar/seminar-papers/31Aug10.pdf

Pawson R (2006) *Evidence-based Policy: A Realist Perspective*. London: SAGE.

Pawson R and Tilly N (1997) *Realist Evaluation* London: SAGE.

Picciotto R (2012) Experimentalism and development evaluation: Will the bubble burst? *Evaluation* 18(2): 213−29.

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

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, 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, 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 CC (1987) *The Comparative Method: Moving Beyond Qualitative and Quantitative Strategies*. Berkeley and Los Angeles: University of California Press.

Ragin CC and Becker H (eds) (1992) *What is a Case? Exploring the Foundations of Social Inquiry*. New York: Cambridge University Press.

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

Rao A and Woolcock M (forthcoming) But how generalizable is that? A framework for assessing the external validity of complex development interventions. World Bank, mimeo.

Ravallion M (2001) Growth, inequality and poverty: looking beyond averages. *World Development* 29(11): 1803−15.

Ravallion M (2009) Should the randomistas rule? *Economists' Voice* 6(2): 1−5.

Rothwell PM (2005) External validity of randomized controlled trials: 'To whom do the results of this trial apply?' *The Lancet* 365: 82−93.

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

Seok J, Warren HS, Cuenca AG et al. (2013) Genomic responses in mouse models poorly mimic human inflammatory diseases. *Proceedings of the National Academy of Sciences* 110(9): 3507−12.

Shaffer P (2011) Against excessive rhetoric in impact assessment: overstating the case for randomised controlled experiments. *Journal of Development Studies* 47(11): 1619–35.

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

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.

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

Yin RK (2009) *Case Study Research: Design and Methods*, 4th edn. Thousand Oaks, CA: SAGE.

Michael Woolcock is Lead Social Development Specialist in the World Bank's Development Research Group, where he has worked since 1998. He is also a (part-time) Lecturer in Public Policy at Harvard University's Kennedy School of Government. An Australian national, he has an MA and PhD in sociology from Brown University.