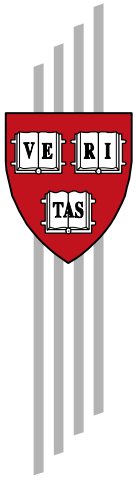
Reasons for Using Mixed Methods in the Evaluation of Complex Projects

Michael Woolcock

CID Faculty Working Paper No. 348
March 2019



Working Papers
Center for International Development at Harvard University

Reasons for Using Mixed Methods in the Evaluation of Complex Projects[1]

Michael Woolcock, World Bank and Harvard Kennedy School

March 2019

Abstract

Evaluations of development projects are conducted to assess their net effectiveness and, by extension, to guide decisions regarding the merits of scaling-up successful projects and/or replicating them elsewhere. The key characteristics of 'complex' interventions – numerous face-to-face interactions, high discretion, imposed obligations, pervasive unknowns – rarely fit neatly into standard evaluation protocols, requiring the deployment of a wider array of research methods, tools and theory. The careful use of such 'mixed methods' approaches is especially important for discerning the conditions under which 'successful' projects of all kinds might be expanded or adopted elsewhere. These claims, and the practical implications to which they give rise, draw on an array of recent evaluations in different sectors in development.

[1] The views expressed in this paper 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 extends and updates work previously published in Woolcock (2009, 2013) and Alcántara and Woolcock (2014). Versions of the central arguments in this paper have been presented at numerous conferences and seminars around the world; I am grateful to attendees at these gatherings for their insight comments and probing questions.

Introduction

In the field of public policy in general – and international development in particular – project evaluations serve two core purposes. The first such purpose is to reach substantive conclusions, on the basis of formal empirical strategies, regarding the nature and extent of the net impact a specific project (or broader portfolio of interventions) has had on targeted populations, e.g., in a particular country (or across a specific sector). Controlling for other factors, did this microfinance project for women in rural Bangladesh reduce poverty?[2] Do participatory programs in Indonesia empower otherwise marginalized groups (such as women) to have a greater influence on collective decision-making?[3] Does using contract teachers in Kenya improve student performance?[4] If the evaluation strategies used to address such questions meet certain professional standards, it is presumed that policymakers and project managers will be in a stronger position to determine whether or not the intervention in question has in fact 'worked'. The more sophisticated the evaluation, the more granular these decisions can be. Has the intervention been more (less) effective for some groups than others? Have particular *aspects* of a given intervention worked more effectively than others? Enhancing the frequency and quality of decisions made on this basis is the essence of widespread calls for taking an "evidence-based approach" to policy (Cartwright and Hardie 2012).

The second core purpose, which is an extension of the first, is to help decision-makers from different contexts draw inferences regarding whether to replicate a demonstrably 'proven' intervention elsewhere, or to scale it up (either to larger numbers of the same target population or to new populations). If a pilot intervention in rural Bolivia seeking to reduce maternal mortality is deemed to have 'worked', should it now be expanded to the cities? Do the "rigorous" positive findings from a deworming project in Kenya warrant its adoption in neighboring Tanzania? What about in Mongolia? Methodologically speaking, the first set of questions pertain to internal validity (or identification) concerns, while the second set to external validity (or generalization and extrapolation).[5] As we shall see, even carefully identified single-method assessments of what I will call 'complex' interventions struggle to address key concerns pertaining to replication and scaling. Appropriately integrated, however, answers to both sets of questions can serve the broader purposes of enhancing 'learning' (so that subsequent decisions regarding a project's design and implementation are made more prudently) and 'accountability' (so that outcomes, such as they are, can be explained on a firm foundation to project recipients, managers, funders, and – if public money is being used – to taxpayers).

This is the conventional way in which evaluation work is framed and discussed, certainly among elite researchers (even if they give vastly more attention to internal validity concerns). Such discussions are necessary and important, and they elicit a range of methodological issues, the resolution of which, as we shall see, is likely to entail using a combination of qualitative and quantitative approaches – i.e., mixed methods. Even so, all such approaches focus largely on

[2] See, for example, Pitt and Khandker (1998). Needless to say, such evaluations invariably elicit criticism (legitimate and otherwise) on both methodological and political grounds (e.g., Roodman and Morduch 2014).
[3] See Mansuri and Rao (2012) for a review of empirical findings (and associated policy claims) from studies from around the world assessing the effectiveness of various 'participatory' development projects.
[4] See Bold et al (2013).
[5] The distinction between internal and external validity (as well as construct validity – the extent to which the specific phrasing of concepts such as 'welfare' in survey instruments accurately reflects how they are understood in everyday life) comes from Cook and Campbell (1979). Construct validity issues are discussed briefly below.

assessing what Goertz and Mahoney (2012) call "the effects of causes": one starts with a given 'cause' (e.g., a program to immunize babies) and then seeks to discern its net effects (e.g., on infant mortality). But many social problems don't (yet) have known solutions, and the most vexing of them are so idiosyncratic that it is highly unlikely that any putative solution deemed to work "there" would also work "here", meaning that considerable adaptation is likely to be required, both upfront and during the implementation process. Such projects are likely to yield widely divergent outcomes across time, place and groups, and as a result require specific explanations for why some places or groups did so much better than others. In such instances, researchers are assessing "the causes of effects": beginning with particular outcomes and then working their way back up the implementation trail to discern when, where and how the critical junctures occurred. Here too, as well shall see, mixed methods approaches are central to generating sound and useable answers.

To narrow our focus somewhat, our concern in this chapter is with such 'complex' projects. In one sense, of course, all policies and projects are far from straightforward, and the methodological challenges outlined above are vexing enough even when it comes to assessing the impacts of relatively 'simple' interventions, such as roads and bridges. For present purposes, such interventions are 'simple' because, for the most part, they are characterized by (a) few ongoing interactions between people being required to realize the intervention's stated objectives (a bridge is inanimate); (b) interactions that do take place leave little room for human discretion (toll collectors perform routine tasks); (c) problems that arise during implementation and maintenance having known (or readily discernable) solutions (fixing potholes, reinforcing girders); and (d) the service performed by the intervention (enhanced connectivity, vastly lower transportation costs) being welcomed by the vast majority of the target population, especially powerful elites.[6]

The very opposite of these four criteria characterize 'complex' interventions such as taxation, justice and social work. For example, if one is implementing a new program to enhance the welfare of children in 'at-risk' households – one which may entail physically removing children from what are deemed to be unsafe family environments – the entire space is characterized by many interacting people, all of whom are exercising considerable discretion, deploying or living with the consequences of a 'solution' whose efficacy is inherently imprecise, doing so in the face of (very likely strong and emotionally wrenching) resistance. What is the ethically sound "rigorous" methodology for assessing the virtues and limits of such a program? Whatever minimally serious evaluation strategy is deployed, its likely finding will be that – befitting the findings of other complex interventions – it worked wonderfully for some, had little effect on others, and was diabolically awful for still others. Even when carefully designed, fully supported (politically and financially), and faithfully implemented, complex interventions are characterized by the highly variable outcomes they generate over time, space and groups – because the intervention's structural characteristics and implementation modality interact with 'contexts' in inherently idiosyncratic ways. By construction one can create a mathematical 'average' impact of such projects, but perhaps the more instructive statistic is the standard

[6] To be sure, the very existence of the bridge, or securing the land needed to make way for the road, may be deeply controversial, but the functional tasks these forms of infrastructure provide – namely, enhancing the ease and speed of travel, and lowering transportation costs – do not themselves provoke coordinated resistance, as does (say) efforts to regulate powerful financial companies. So, to be more precise, there may well be 'complex' *aspects* of standard infrastructure projects (such as peaceably securing the land on which they will reside).

deviation – the variability around the average that, if carefully monitored over time, can be a fruitful basis of iterative learning. This monitoring itself, however, and accurately discerning the 'lessons' from it, will require access to a broad array of theory and methods.

The central premise of this chapter is that complex interventions, as defined above, are best assessed by 'mixed methods' – i.e., an array of integrated qualitative and quantitative approaches to research design, measurement, analysis and interpretation that exploits the comparative advantage of each approach in the joint pursuit of knowledge enabling real-time adjustments. Complex projects "learn" in a manner to which human learn complex tasks such as speaking a foreign language or playing a musical instrument: by extended trial and error. In the sections that follow, the strengths and weaknesses of stand-alone approaches to evaluation are outlined, along with a discussion of the importance of embedding empirical findings regarding project impacts in a theory of change that accommodates the likely trajectory of that impact over time. (Most of the examples come from international development, but they have been chosen because of the broader applicability of the common underlying principles.) Such analyses form the basis of a third section exploring the conditions under which empirical claims about the impact of a given complex intervention might be generalized to novel contexts, scales of operation and implementing agencies. A concluding section reflects briefly on the rising and expanding role for complex policy interventions, and the corresponding demand this will place on evaluators to become adept at assembling interdisciplinary teams (since it is unrealistic to expect any single evaluator to be fully competent in all methodological approaches).

The Complementary Strengths and Weaknesses of Different Methodological Approaches

Research and evaluation methods in the social sciences are typically categorized as either quantitative or qualitative, as are the data that these respective methods deploy (Hentschel 1999).[7] Quantitative methods, such as econometrics, use large amounts of numerical data derived from primary (e.g., household surveys) or secondary (government records) sources to draw inferences regarding relationships between categorical variables (e.g., age, occupation, income, health). Since it is rare to obtain such data on entire populations, careful attention is given to sampling concerns and specifying the confidence one has in both the strength of the measured relationships (net of other factors, such as non-random selection into groups) and the conditions under which these relationships might hold for the larger population. Largely because of this capacity to speak to trends and relationships in large populations, quantitative methods and data assume a privileged status in public policy deliberations. Qualitative methods, such as those of mainstream anthropology, focus on understanding the intricate details of the processes and meanings associated with social interactions within and between particular groups. As such, qualitative methods (interviews, observations, textual analysis) tend to be associated with qualitative data (words, images)[8]; less concern is given to demonstrating whether emergent

[7] More nuanced distinctions include comparative methods as a separate third epistemological approach (e.g., Ragin 2014) but the use of such methods is relatively rare in project evaluation and thus are not discussed here.

[8] A benefit of distinguishing between methods and data is that it creates a space for recognizing that quantitative methods can be used on qualitative data (e.g., assessing the frequency of certain words in books or newspapers over hundreds of years, as search engines now make possible) and that qualitative methods can be used to generate quantitative data (e.g., when medical anthropologists collect data on the height and weight of children in remote villages as a guide to assessing their overall health status). Qualitative methods (such as 'anchoring vignettes'; see Hopkins and King 2010) can also be used to enhance inter-rater reliability in response to subjective questions in

findings (e.g., from a single village) are 'representative' of the larger population from which they are drawn (e.g., a region or country) since such claims are rarely made or expected. Qualitative methods are especially useful when the interventions to be evaluated increase in complexity (i.e., require many discretionary and face-to-face transactions, and are contentious[9]), when the 'context' itself is highly variable (and perhaps volatile), when the quality and availability of existing data is poor, and when insights are sought on specific types of impacts on specific groups (e.g., the effectiveness of a project for ethnic minorities, informal firms or illegal immigrants, who may not be adequately represented in formal surveys). Qualitative methods can also be useful when evaluating small-N interventions such as regulatory reforms at the national level, or automation of procedures in one single agency.[10]

For the purposes of understanding the impact and generalizability of claims pertaining to complex projects, perhaps the simplest but most fruitful distinction between these quantitative and qualitative approaches is to argue that the former focus on 'breadth' where the latter focus on 'depth'. The main rationale for the systematic integration of qualitative and quantitative methods in the evaluation of projects (of any kind) is that both approaches complement the others' limitations; this is particularly so with regards to the 'breadth' and 'depth' of information that together is needed to optimally describe and explain outcomes stemming from complex phenomenon. In this way, integrating qualitative methods in impact evaluation helps reveal the ways in which different causal mechanisms—singularly or in combination—generate observed outcomes and thereby enable evaluators to assess the intervention's broader theory of change[11]; i.e., both whether and *how* impact is achieved in a specific instance, and the conditions under which this impact might be expected elsewhere or at larger scales of operation (Bamberger et al 2010, Clark and Baidee 2010). Table 1 summarizes the key ways in which both methodological approaches are used in the collection, design, analysis and interpretation of data in project evaluations.

Table 1: Characteristics of Quantitative and Qualitative Methods in Project Evaluations

large-scale surveys. Space precludes exploring these particular types of approaches in this chapter, since they are the exception rather than the rule in terms of how most evaluations of complex projects are conducted.

[9] Thus delivering the mail is a 'simple' (logistical) task while promoting women's empowerment in rural Pakistan, or regulating powerful companies, is a highly 'complex' one (see Andrews et al 2017).

[10] Small-N cases are those in which insufficient units are available to be assigned to comparison groups to get the sufficient statistical power to run an experimental or quasi-experimental design. For a helpful discussion on this point, and how concerns surrounding it might be addressed, see Ruzzene (2012).

[11] 'Mechanisms' here refers to specific processes causally connecting discrete variables; 'normal science' advances when these processes are understood ever more precisely and at smaller units of analysis. (A canonical example is the refinement of knowledge from 'citrus fruits' to 'Vitamin C' as the *mechanism* responsible for alleviating scurvy among sailors.) Strictly speaking, a true mechanism is time and context invariant – taking Vitamin C will always and everywhere reduce the likelihood of scurvy – though relatively few of these have been identified in the social sciences (for reasons partially articulated in Henrich et al 2010). A 'theory of change', on the other hand, is a broad (aspirational) statement asserting, on the basis of logic and reason, how the provision of certain inputs (e.g., cash given to poor households) will, through a long administrative implementation chain in a particular context, lead to outputs (increased school attendance) that, in turn, generate a desired policy outcome (e.g., enhanced learning) and impact (higher income, reduced poverty). A given intervention can 'fail' because of breakdowns at any point along this implementation chain, which is why a comprehensive *theory of change* needs to be specified from the outset – the better to anticipate where such breakdowns might occur, and to respond accordingly.

	Quantitative Methods	Qualitative Methods
Research Questions	Usually derived deductively (e.g., from knowledge gaps in the literature); seek to demonstrate 'precise' causal effect (impact) of x on y for relatively large populations; can also draw on qualitative insights to refine/adapt questions for specific contexts	Usually derived inductively (e.g., by refining questions as they emerge in situ); focus on process concerns—how outcomes were attained, how different types and combinations of mechanisms generated different outcomes for different groups
Data Collection	Use data collection methods such as surveys with closed ended questions; this standardizes but limits the depth and variability of the information that is obtained	Use data collection methods such as focus groups to capture in-depth, context-specific information; also used to ensure that questions in surveys are worded and sequenced in ways that all parties understand ('construct validity')
Evaluation Design	Seeks to reduce selection bias (and other confounding factors), and to ensure representativeness and comparability of project and non-project samples to enhance quality of statistical inference ('internal validity')	Can help to discern and discuss issues that are 'unobservable' statistically (including identifying good instruments); weaknesses in 'breadth' and representativeness are compensated for by strengths in 'depth' and understanding of causal mechanisms
Analysis and Interpretation	Quantifies the magnitude of impact to try to determine *whether* an observed outcome can be causally attributed (probabilistically) to the intervention; but even the most 'rigorous' ('well-identified') analysis rarely provides warrant for inferring that similar results will obtain elsewhere (or at larger scale) ('external validity')	Is best suited to informing discussions regarding *how*, *why* and *for whom* a given intervention worked (or not); thus can help explain (and foster learning from) variation in outcomes and/or implementation processes, and usefully contribute to discussions about the possible generalizability of given findings to novel contexts, populations and scales of operation

Source: Alcántara and Woolcock (2014)

Another benefit of using qualitative and mixed methods in project evaluations is that they can enhance the robustness of the underlying model of causal inference (i.e., improve internal validity) and thereby diminish the influence of various sources of bias (e.g., selection bias, by observing 'unobservable' factors shaping program placement and participation) and measurement error (e.g., discrepancies in terms of how survey questions are understood by

respondents and researchers).[12] Results obtained from qualitative analysis may support the conclusions obtained from the quantitative research but enable researchers to go beyond the measurement of impacts and provide specific evidence of *how* impact was achieved and for whom – i.e., it can facilitate the exploration of variation across time, space and groups, by showing how local context characteristics and implementation dynamics interact. In a recent study of a national community development project in Indonesia, for example, even neighboring villages performed quite differently; a key factor shaping this variation was whether local leaders supported or resisted the project, even though these villages were participating in the same project being implemented by the same people (Barron et al 2011).

In other instances, however, qualitative research might qualify or even contradict the findings emerging from quantitative approaches, in which case the research team needs to work together to resolve the anomalies; these deliberations, if done carefully, can serve to enhance the confidence the project team (and stakeholders in the reform process, including policy makers) has in the final conclusions and the policy implications to which they give rise (Woolcock 2009; Rugh et al 2011). Results from a quantitative evaluation of a jobs program, for example, may show that wages significantly increased for program participants, and thus conclude that it was a success, while a qualitative assessment may find that program participants reported heightened levels of stress and health problems, and thus conclude that the program was a failure. Which interpretation is correct? Combining both findings may lead to a more nuanced and helpful conclusion, namely that real wage increases were achieved but at the price of considerable welfare declines for certain groups, enabling corresponding adjustments to be made in subsequent iterations of the program. Even when the empirical findings derived from different methods align, an iterative dialogue between qualitative and quantitative perspectives can contribute to a more comprehensive interpretation of the results – what they mean, and what their implications are for policy and practice (Shaffer 2013).

In short, the systematic combination of quantitative and qualitative methods helps evaluators to optimize the likelihood that their findings (and interpretations of those findings) will lead to accurate inferences about the effectiveness of interventions, and how this effectiveness varies across time, contexts and target groups. It achieves this primarily by using the strengths of one approach to offset the weaknesses of the other (Rao 2002; Rao and Woolcock 2003). Other instances where quantitative and qualitative methods can be combined in the evaluation process include:

- Generating hypotheses about an intervention's effectiveness from theory, experience and qualitative research and then testing their ability to be generalized with quantitative techniques.
- Identifying contextual factors, processes and causal mechanisms via qualitative methods and assessing them further via quantitative methods (e.g., Ludwig et al 2011) and/or additional qualitative analysis.

[12] Quasi-experimental designs, for example, present the risk of selection bias due to unobservable factors that affect participation and outcomes which are neither easy to measure, are not known by the researcher or are time variant. Using qualitative methods enables researchers to identify potential instrumental variables or identify those time variant and invariant unobservable variables.

- Applying quantitative sampling techniques to units of qualitative data collection, and/or findings from qualitative analysis and using them to inform the design of quantitative data collection tools (i.e., household or firm surveys).
- Using qualitative findings to see if they support, explain, qualify or refute quantitative findings regarding an intervention's impact (Rao et al 2017).

I address these and related issues in more detail below.

Even though the deployment of mixed-method approaches has been increasing in economic development impact evaluations, most notably in health, to date relatively few impact evaluations can be identified as truly using a mixed-method approach. For example, only three percent of 3ie's portfolio has used a mixed-method approach[13], and neither J-PAL nor World Bank databases formally record whether mixed methods were used in a given evaluation. In the following sections we provide some examples of how qualitative methods have been deployed in each stage of the standard evaluation cycle. Although these studies did not use a systematic integration of methods, they are useful to showcase the fruitfulness of deploying mixed methods in specific stages of the evaluation. It bears repeating that, ideally, the most valid and useful findings are likely to emerge when both qualitative and quantitative methods can be integrated at different stages, enabling their systematic combination to exploit the strengths (and minimize the weaknesses) of using one method alone.

Understanding Impact Trajectories

Any hypotheses or claims about change processes must incorporate time (by when it is reasonable to expect that a net impact will be attained – six months, six years?) and the high likelihood that the trajectory of that change will be non-linear (e.g., a J-curve or step function). Giving inadequate attention to changing circumstances and the possibility of non-linear impact trajectories can lead to claims about impact that turn out to be premature, thereby forming an inaccurate basis for future projections. For example, a study that evaluated the impact of an export promotion-matching grant for small and medium-sized enterprises (SMEs) in Tunisia found that in the short term, beneficiary firms showed higher export growth and export diversification than those of the control group. However, in a subsequent study it was found that the effects were not sustained over time, an issue that the authors highlighted as commonly overlooked in the literature (Cadot et al 2012). The authors of the follow-up study even mention that these types of reforms have not been explored in the long term, questioning the sustainability of what in the short-term was found to be "successful" (Cadot et al 2012). Ravallion (2009) warns that the assessment of short-term impacts is common in impact evaluation, generating a "myopia bias" that can lead not only to erroneous conclusions but also to decisions to scale-up policies and programs without knowing the underlying factors of impact that can lead to negative spillovers.

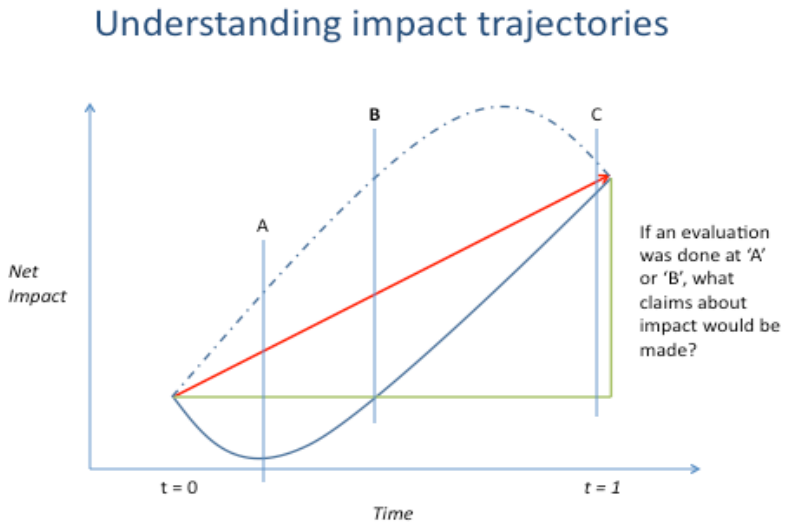
[13] Better Evaluation Blog, August 2013. "Mixed methods in evaluation Part 3: Enough pick and mix; time for some standards on mixing methods in impact evaluation." http://betterevaluation.org/blog/mixed-methods-part-3

Box 1: How Impact Trajectories Shape Interpretations of Impact

Four months after planting, we do not conclude that the growth of oak trees (which takes years) is 'less effective' than the growth of sunflowers (which takes weeks) because science and experience tell us what it is reasonable to expect by when. The same logic should apply to development interventions. The important implication is that when assessing an intervention at two points in time, evaluators must have (or build) a solid theory of change – on the basis of experience, evidence or theory – to specify the mechanisms (processes) by which they expect given inputs to generate observed outcomes, and over what time-frame and trajectory it is reasonable for these outcomes to emerge (Woolcock 2009, 2013). Both qualitative and quantitative methods are needed to do this well. (Most complex to assess of all, of course, are those interventions that have no consistent impact trajectory.)

A central issue for both causal inference and policy extrapolation is that methods per se, no matter how 'rigorously' and comprehensively they are applied, do not on their own provide a clear basis for discerning whether an intervention is working or is likely to do so in the future; for that, the empirical findings must be guided by theory and experience. Put differently, *the implications of evidence are never self-evident.*

Consider the figure below, which exemplifies four different impact trajectories and three different points in time at which an evaluation could be conducted: without knowledge of the likely impact trajectory associated with a given intervention (say, roads versus schools versus immunization versus land titling), and thus knowledge of what it is reasonable to expect by when, wildly inaccurate conclusions regarding the intervention's efficacy could be drawn. If the intervention was evaluated at point C, a fortuitously consistent story would emerge since all four trajectories converge on a similar net impact between 'baseline' (t=0) and follow-up (t=1). (And the timing of the follow-up is largely determined by political and administrative imperatives, not scientific ones.) But if the intervention was evaluated at point A, four very different conclusions regarding the intervention's net impact – ranging from spectacular success to dismal failure – would be drawn, even if the intervention was being assessed via an RCT. The shape of the trajectories, when extended into the future, has correspondingly important implications for the claims we make about the intervention's likely impacts down the road.



Source: Woolcock (2013)

This dynamic can be seen in a World Bank-supported land reform project in Cambodia, which was hailed (rightly) as an initial success. But a mismatch between the reform's expectations and the capacity of the administrative system to implement them on a larger scale, especially in sensitive peri-urban areas, generated stress on the demand side and weakened (in fact almost collapsed) the capacity of the system (Adler et al 2008, Biddulph 2014). Hence, generating in-depth contextual information is key to identifying the factors that are shaping the nature and extent of an intervention's impact trajectory (see Box 1), and to sustaining a commitment to equitably negotiating those aspects of implementation that may be contentious. Such information also plays a key role in decisions about whether, when, where and how the intervention might be scaled up (or shut down, for that matter).

Hence, one important question that arises is: *when* should impacts be measured? By using qualitative methods to understand the context and by drawing on a range of experiences elsewhere, evaluators can derive informed knowledge of the change process (whether it be initiated by firms, governments, NGOs or others), and thus help to more accurately specify what outcomes the intervention can be expected to generate over a given timeframe. Failure to do so can lead to claims about impact that are accurate only at a certain (often arbitrary) time period, when a fuller rendering of the path taken so far, and the path(s) that is likely in the future, is needed to guide decision-making.

Integrating Qualitative and Quantitative Methods into 'Complex' Project Evaluations

As previewed above, qualitative analysis and data collection can complement quantitative techniques in the evaluation design to address common challenges such as identification (i.e., inference regarding causal relations), construct validity (assuring the quality of data itself) and model specification, but are also crucial for understanding the role of implementation quality and 'context', and interpreting extant empirical findings. These latter issues are especially salient in the evaluation of 'complex' projects, such as those pertaining to governance and legal reform. I address these issues in turn:

(a) Identification: Qualitative data collection and analysis can be helpful in informing and selecting samples (whether of people, places or issues) of interest. For example, in-depth interviews or focus groups might be used to identify firms or individuals with "entrepreneurial" behavior, or to identify what constitutes entrepreneurial behavior according to the context and prevailing social norms. Once firms are identified, quantitative methods can be applied to the population of interest to make the sample (more) representative. Another common identification strategy is selecting samples (or even stratified samples) of interest from the sampling list with specific characteristics; qualitative research can then be conducted on those selected individuals or units of interest to help explain common or different characteristics, or to explain variance or outlier behavior (Tedlie and Yu 2007). This technique is particularly useful when sample sizes are small.

Qualitative data collection methods have also been useful in refining the identification strategy and diminishing the risk of selection bias, especially for quasi-experimental studies where it is difficult to control for unobservable variables. An example is Bloom et al (2013), who assessed the impact of management practices in firms' performance in India by conducting retrospective interviews and observation assessments at the factories of a representative sample

of firms. Data gathered was used to confirm that there was no significant difference between the project and non-project firms. This study shows the importance of integrating both qualitative and quantitative sampling techniques to obtain a sample representative of the population with the specific desired characteristics. Such quantitative sampling techniques help to ensure that qualitative samples are adequately representative, and contribute to ensuring that claims regarding the implications of these findings for wider populations are well founded.

(b) Construct Validity: Qualitative analysis may be useful to explore the dimensions of the indicators used in the design. Definitions of concepts such as 'corruption', 'justice' or 'transparency' may vary widely across individuals, locations or sectors. Exploring the meanings of indicators according to the context and incorporating them into the quantitative data collection methods is not only critical to obtain accurate data from surveys, but also plays a key role in establishing and explaining causation.

As an example, the concept of 'delay' in clearing goods in a border post may have different meanings depending on the sector and for people working at different points in the distribution channel. For importers of ultra-fresh products, a 'delay' might be understood as more than one day, while for other sectors (e.g., processed food), it might represent more than three days. The definition may vary per location. A mixed-method approach can contribute to incorporating different dimensions of particular indicators (Shaffer 2013). Another example is the concept of 'human welfare'. The most widely assessed measure of human welfare may be income (or expenditure), but if this was the only indicator chosen to assess a project's effectiveness at improving 'human welfare', it would be considerably inadequate as a basis for an empirical conclusion if that conclusion had not been informed by insights from qualitative research potentially showing (say) that income gains were indeed attained, but at the price of increased stress and deteriorating mental health. Similarly, standard quantitative measurements of poverty such as consumption per capita can be weighted according to local or contextual definitions or perceptions of what 'poverty' means (Kristjanson et al 2010).

Understanding the dimensions of the indicators in their context (and for different personnel within a given context) is necessary to understand what is intended to be measured. Rao (2002) describes how a survey on the incidence of domestic violence in India generated rates far below expectations. Initial survey results suggested that the incidence of household violence in India was even lower than in the US, but when researchers conducted qualitative analysis of this issue they found that domestic violence was understood differently relative to the context (e.g., a slap would not be considered as domestic violence by the average Indian household). Hence, the survey questions and results were inaccurate and did not reflect an accurate domestic violence situation. Even though quantitative approaches can be applied to measure changes in these outcomes, understanding the definitions of concepts as understood by different respondents is key to establishing valid quantitative measures for these concepts (i.e., to ensuring high 'construct validity').

(c) Causation and model specification: By having detailed knowledge of a particular context, qualitative work can be helpful in solving endogeneity problems[14] and can reveal the direction of

[14] In evaluation, endogeneity problems stem from biased estimates of impact due to issues such as omitted variables or measurement error, which weaken the claims of attribution. In principle, experimental designs greatly reduce these problems by ensuring that any such biases are at least equally present in the treatment and control groups.

causality by identifying instrumental variables (Ravallion 2000; Rao and Woolcock 2003). (Some qualitative researchers also argue that techniques such as process tracing can be used to make causal claims of their own – e.g., Bennett (2010) – and note that case study evidence is routinely the basis on which causal arguments are made and defended in 'real world' settings such as court rooms – Honore (2010), Cartwright (2016). However, I shall not address the details of such matters here.)

(d) Quality and reliability of data collection: Understanding the context through qualitative analysis is not only useful with regards to knowing what should be assessed in a survey or what should be included in an equation. It also contributes insights as to how and to whom questions should be asked or assessed, given that the quality of the data obtained depends on the collection methods used with specific objectives in specific contexts. As an example, Sana et al. (2012) conducted a study in the Dominican Republic and found that respondents answered differently depending on the type of questions asked by type of interviewer (local or external). They found that respondents reported higher income and higher tolerance towards marginalized groups to external interviewers compared to the responses given to local interviewers. Hence, qualitative methods can help to improve the quality of the data by exploring the best ways in which a question should be asked, how and to whom it should be asked, and by whom.

Parallel qualitative data collection techniques such as those generated by participant observation or case studies can also help to assess the reliability and quality of the data collected through surveys. The IFC Lima Tracer Study, for example, which assessed the impact of firm formalization on the performance of micro firms in Lima, found significant divergence from survey responses when the team conducted in-depth interviews to try to understand the low demand for operating licenses. Researchers explain that this may happen because "questions involving a moral issue, such as complying with the law, tend to be answered 'correctly', but not necessarily honestly" (Alcazar and Jaramillo 2011).

(e) Implementation factors. Qualitative data collection and analysis generally ask and answer different questions from quantitative approaches (when the aim of the mixed-method approach is not triangulation), in the process uncovering other factors that may be shaping observed impacts such as the institutional framework (i.e., formal laws and regulations, and informal customs and norms). Contextual analysis contributes to assessing the institutional capacity of local agencies involved in the project (i.e., financial resources, political support, power of implementation) the political economy and the forces supporting (or undermining) the reform, and so on. These factors, which are difficult to measure quantitatively, may influence the quality of implementation and outcomes/impact. Qualitative data collection assessing the process of implementation can provide insights of how and why outcomes and impact were achieved. One criticism of conventional impact evaluations is that when expected impacts are not found, given that there is a lack of process evaluation or monitoring, it cannot be inferred if the absence of impact was because of failure of the design/causal link or the failure of implementation (Bamberger et al 2010; Rao et al 2017).

Qualitative methods can be especially useful with regards to assessing the process and quality of implementation. For example, in implementing competition reforms, it has been found that larger impact in selected outcomes is achieved when effective enforcement is implemented (Kitzmuller and Licetti 2012). The implementation of effective enforcement could be analyzed

starting from the political context through the analysis of secondary data such as newspapers, by conducting direct observation or process tracing.[15] Results obtained from qualitative data collection methods may be transformed to variables that reflect these issues and can be incorporated into the econometric study, or they can be used in parallel to explain quantitative results.

(f) Data analysis and interpretation. Qualitative analysis can contribute to internal validity by verifying the connections between the causal mechanisms identified in a quantitative analysis. (Similarly, if its findings are contradictory, it may provide an alternative explanation or lead to further research.) As an example, in an evaluation assessing the demand for formalization among firms in Sri Lanka, researchers wondered if the large shifts in profits that few firms reported were attributed to formalization or were due to measurement error (De Mel et al 2013). The researchers conducted case studies to ensure that the findings were not driven by measurement error and to articulate the mediating channels through which formalization helped the firms that benefitted most. The qualitative analysis supported the quantitative findings and confirmed the causal mechanisms demonstrating that formalization led to increased firm profits. The qualitative analysis shed light on how formalization helped firms by allowing them to issue receipts and thereby become suppliers in larger value chains – in a very effective way.

Another relevant example is again the Lima Tracer Study, in which researchers used as baseline data firms operating without a license and used incentives such as fee waivers for the treatment group. The analysis found no significant impact on outcome variables. In addition, it was noted that firms were not eager to take the incentives. Through a qualitative study applied to a smaller sample, it was possible to distinguish behavioral characteristics of entrepreneurs associated with license acquisition. Information obtained through in-depth interviews revealed that there are two distinct groups among the entrepreneurs—"typical entrepreneurs" and "survival entrepreneurs" —and that this distinction may be considered a determinant in the decision to obtain a license. In addition, managers from micro firms did not perceive important benefits from formalization and recognized that the cost of the license is a real barrier for the formalization process, but not the most important. These interviews led to the conclusion that, in fact, there is not a high demand for operating licenses, an issue that was not captured through surveys, which also explains the low take-up and impact obtained.

The qualitative analysis was not initially contemplated; the original design was mainly a quantitative approach. As many companies did not accept the incentives, the research institute (GRADE) decided to conduct an in-depth study with a qualitative focus. Given the insightful findings obtained from the qualitative analysis, GRADE started using mixed methods in its impact evaluations. The most common design now used is to initially conduct a qualitative study to understand the context and develop the questions for surveys and find insights regarding the outcome variables that should be taken into account. After the quantitative analysis is conducted, a second qualitative analysis is used to explain or dig deeper into the results found.[16]

[15] Process tracing is a tool of qualitative analysis that contributes to drawing descriptive and causal inferences from diagnostic observations undertaken chronologically (Collier 2011).
[16] Information obtained from a telephone interview with Lorena Alcazar, November 2012.

Box 2: An Example of Mixed-Methods Evaluation of a Complex Intervention

One example that illustrates the iterative systematic approach is an assessment of the Kecamatan Development Project (KDP, a national community-driven development program) in Indonesia on local conflict dynamics (Barron et al 2011). KDP's objective was to provide block grants to local communities, who would then allocate this money to those projects community members themselves deemed most pro-poor, sustainable and cost-effective. This allocation process took place in community forums, but not every proposal was funded, generating the potential for conflict if villagers perceived that outcomes were a function of non-merit-based procedures (or worse). The evaluation's objective was to assess whether and how these forums improved local governance; the hypothesis was that participating in KDP creates robust civic spaces and deliberative skills which enable local conflicts to be constructively addressed. One major challenge was that 'conflict' is notoriously hard to measure, and what little data there was had been collected from village leaders (who had obvious incentives to under-report the incidence of conflict on their watch). A mixed-method approach was used to find a novel way to measure conflict (which included a comprehensive analysis of local newspapers) and the mechanisms by which it is initiated or resolved (discerned via key informant interviews). In addition, it was critical for the evaluation to understand the causal chain of events, which was only possible with a deep qualitative analysis (which was generated by collecting dozens of cases of conflict pathways in program and comparable non-program villages).

An iterative strategy for integrating the quantitative and qualitative analysis was used. An initial period of qualitative fieldwork was pursued for three months. The villages were selected using a quantitative sampling frame (using propensity score matching (PSM) techniques derived from nationally representative household surveys), but the final selection of the best match of program and non-program villages was made using detailed contextual knowledge (since a well-understood weakness of PSM is that it only matches on 'observable' characteristics). This was critical to capture heterogeneity of the population and increase the validity of the results. This initial work contributed to the sampling of districts, research hypothesis formulation and design of adequate survey questions. Once the identification of a "counterfactual" was done using qualitative analysis and supported by quantitative methods, data was collected from a survey administered to a larger sample of households and used to assess the generality of the hypotheses emerging from the qualitative work. In addition to the quantitative analysis, the analysis of case studies of local conflict, interviews, surveys, key informant questionnaires and secondary data sources as newspaper evidence, provided a broad range of evidence to assess the validity of the hypotheses stating the conditions under which KDP could (and could not) contribute to solve local conflict.

Another common situation in which the usefulness of mixed methods can be seen is small-N evaluations, such as the introduction of a business regulatory reform at the national or sub national level. Such reforms, by their very nature, make the construction of a counterfactual difficult or even impossible. In such circumstances, a process of elimination can be deployed to systematically identify and rule out alternative causal explanations of observed results. For example, firm performance could be attributed to the improvement of the business climate but this could be happening in ways unrelated to the actual business entry reform, such as via improvements in infrastructure or more information being available on business opportunities. A thorough qualitative analysis of the processes by which positive outcomes were attained could enable one to establish a detailed causal chain and define how the specific context interacts with the reform and outcomes. Quantitative approaches can be used in parallel for triangulation purposes, or can contribute by helping evaluators avoid some of the typical biases associated with qualitative analysis (such as selection bias), including selecting firms for in-depth analysis using randomization or purposive sampling techniques.

Assessing the External Validity of Complex Interventions[17]

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 'those guys' 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" (Cartwright and Hardie 2012: 137), 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[18]: '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'[19]

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 and Callighan 2013; Byrne 2013), but for present purposes I will create some basic

[17] This section draws on Woolcock (2013).

[18] 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.

[19] 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.

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[20]) 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 the effects of textbooks, de-worming pills, 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 behavior – see baby, insert needle – on the part of front-line agents. 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 behavior 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.[21]

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

[20] See Ludwig et al (2011) for a discussion of the virtues of conducting delineated 'mechanism experiments' within otherwise large social policy interventions.

[21] 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.

in the face of the HIV/AIDS epidemic, social work – 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.[22]

By the definition of complexity offered in this chapter's introduction, 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.

'Implementation Capability'

As noted in the preceding section, 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) program 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

[22] 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.
[23] In more vernacular language we might characterize such problems as 'wicked' (after Churchman 1967); see also Andrews et al (2017).
[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 programs on education outcomes.

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 (Andrews et al 2017) 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 (2014) 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 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% 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.

'Reasoned Expectations'

As discussed above, complex interventions are highly likely to unfold along non-linear trajectories. Accordingly, 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. With variable time frames and non-linear impact trajectories, vastly different accounts can be provided of whether a given project is "working" or not.

[25] Indeed, for a time the high-profile advertising slogan of a large, private international parcel service was: 'We love logistics'.

[26] For a broader conceptual and empirical discussion of the evolving organizational capabilities of developing countries see Andrews et al (2017).

A study 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' 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). 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.

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. 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 (if x gets us y, will 10x get us 10y?), context and implementation capability. Bruhn and McKenzie (2013), for example, show that a business registration program in Brazil that worked wonderfully as a pilot failed as a national project, because at scale citizens perceived it to be a surveillance tool designed by an overbearing state to monitor their business activities. Bold et al (2013) show that an intervention (using contract teachers in schools) that worked well in Kenya when implemented by an NGO was unable to generate the same result when exactly the same intervention was implemented by the government of Kenya.

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. Incorporating an array of complementary theory and methods best suited to addressing these concerns into the evaluation's design and conduct offers the most promising path to more satisfactory inferences and extrapolations. Causal density, implementation capability, reasoned expectations together 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

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

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) (see Widner, Woolcock and Nieto, forthcoming).

4. Conclusion

A defining characteristic of complex development interventions is that – even when carefully designed, politically supported and faithfully implemented – they generate highly variable impacts across contexts, populations and time. A second defining feature is that it is impossible to fully anticipate, up front, all the contingent events and decisions that will need to be made during implementation, meaning that learning in real time from this variation is itself necessary to ensure that positive impacts on target populations are maximized.[28] Discerning this variation, the sources of it, the reasons for it and the implications from it, cannot be done using a singular method (no matter how putatively 'rigorous') or the tools of a singular discipline; of necessity it requires instead the deployment of a mixed methods approach.

From this standpoint, efforts to enhance development effectiveness through evidence derived from project evaluation need to move beyond debates pertaining to the 'rigor' of isolated methods to more concerted attempts to understanding mechanisms driving impact trajectories over time, in different places, at different scales, and in accordance with how well they are implemented. Knowledge of exactly how, where, when and for whom this variance manifests itself is crucial for making accurate empirical evaluations of project/policy effectiveness. Doing this well requires, in the first instance, familiarity with the serious challenges associated with assessing complex interventions and awareness of the *array* of methods that exist to deal with them. It also requires a capacity to discern and to combine, and to work constructively in teams (since, given the degree of specialized knowledge required, it is unrealistic to expect a single person to be fully conversant across these different methodological domains).

Acquiring the knowledge necessary to assess complex interventions will not be a product of simply deploying what some deem to be 'gold standard' evaluation protocols per se, but rather deep engagement with the contexts and processes within which all projects are embedded, and calling upon the full arsenal of research tools (qualitative, quantitative, and comparative-historical) available to social scientists. The future will surely be more rather than less 'complex'; evaluations of interventions addressing these issues must themselves be designed accordingly, rather than imagining that singular approaches can elicit the "key facts" they were not designed to elicit.

[28] Kauffman (2016: xiv) argues that such characteristics render the state of an emergent phenomena 'unprestate-able' – an inelegant but technically accurate description. In these instances, he argues, "[n]ot only do we not know what *will* happen, we often do even know what *can* happen. If we cannot prestate what *can* happen, we cannot know what can happen and thus cannot reason about it. But we must live forward anyway…" (emphasis in original)

References

Adler, Daniel, Doug Porter and Michael Woolcock (2008) 'Legal Pluralism and Equity: Some Reflections on Land Reform in Cambodia' Washington, DC: World Bank. Available at http://siteresources.worldbank.org/INTJUSFORPOOR/Resources/J4PBriefingNoteVolume2Issue2.pdf

Alcántara, Alejandra Mendoza and Michael Woolcock (2014) 'Integrating Qualitative Methods into Investment Climate Impact Evaluations' Washington, DC: World Bank Policy Research Working Paper No. 7145

Alcázar Lorena and Miguel Jaramillo (2011) 'Panel /Tracer Study on the Impact of Business Facilitation Processes on Microenterprises and Identification of Priorities for Future Business Enabling Environment Projects in Lima, Peru' GRADE

Andrews, Matt, Lant Pritchett and Michael Woolcock (2017) *Building State Capability: Evidence, Analysis, Action* New York: Oxford University Press

Baird, Sarah, Francisco Ferreira, Berk Özler, and Michael Woolcock (2013) 'Conditional, Unconditional and Everything in Between: A Systematic Review of the Effects of Cash Transfer Programmes on Schooling Outcomes' *Journal of Development Effectiveness* 6(1): 1-43

Bamberger, Michael, Vijayendra Rao and Michael Woolcock (2010) 'Using Mixed Methods in Monitoring and Evaluation: Experiences from International Development', in Abbas Tashakkori and Charles Teddlie (eds.) *Handbook of Mixed Methods in Social and Behavioral Research* (2nd revised edition) Thousand Oaks, CA: Sage Publications, pp. 613-641

Barron, Patrick, Rachael Diprose and Michael Woolcock (2011) *Contesting Development: Participatory Projects and Local Conflict Dynamics in Indonesia* New Haven: Yale University Press

Bennet, Andrew (2010) 'Process Tracing and Causal Inference', in Henry Brady and David Collier (eds.) *Rethinking Social Inquiry* (2nd edition) Lanham, MD: Rowman and Littlefield

Biddulph, Robin (2014) "Cambodia's Land Management and Administration project" WIDER Working Paper No. 2014/086. Helsinki: UNU-WIDER

Bloom, Nicholas, Benn Eifert, Aprajit Mahajan, David McKenzie and John Roberts (2013) 'Does Management Matter? Evidence from India" *Quarterly Journal of Economics* 128(1): 1-51

Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, Alice Ng'ang'a and Justin Sandefur (2013) 'Scaling-Up What Works: Experimental Evidence on External Validity in Kenyan Education' Washington: Center for Global Development, Working Paper No. 321

Byrne, David (2013) 'Evaluating Complex Social Interventions in a Complex World' *Evaluation* 19(3): 217-228

Byrne, David and Gillian Callighan (2013) *Complexity Theory and the Social Sciences: The State of the Art* London: Routledge

Cadot, Olivier, Ana M. Fernandes, Julien Gourdon and Aaditya Mattoo. (2012) 'Are the Benefits of Export Support Durable? Evidence from Tunisia." Washington, DC: World Bank. https://openknowledge.worldbank.org/handle/10986/12189

Cartwright, Nancy (2016) 'How to Learn About Causes in the Single Case'. Paper prepared for Princeton - World Bank conference on 'The Case for Case Studies: Integrating Scholarship and Practice in International Development'.

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

Casey, Katherine, Rachel Glennerster and Edward Miguel (2012) 'Reshaping Institutions: Evidence on Aid Impacts Using a Pre-Analysis Plan' *Quarterly Journal of Economics* 127(4): 1755-1812

Chong, Alberto, Rafael La Porta, Florencio Lopez-de-Silanes and Andrei Shleifer (2014) 'Letter Grading Government Efficiency' *Journal of the European Economic Association* 12(2): 277-299

Churchman, C. West (1967) 'Wicked Problems' *Management Science* 14(4): 141-142

Clark, Vicki Plano and Manijeh Baidee (2010) 'Research Questions in Mixed Methods Research', in Abbas Tashakkori and Charles Teddlie (eds.) *Handbook of Mixed Methods in Social and Behavioral Research* (2nd revised edition) Thousand Oaks, CA: Sage Publications, pp. 275-304

Collier, David (2011) 'Understanding Process Tracing' *PS: Political Science and Politics* 44(4): 823-30

Cook, Thomas D. and Donald T. Campbell (1979) *Quasi-Experimentation: Design and Analysis Issues for Field Settings* Boston: Houghton Mifflin Company

De Mel, Suresh, David McKenzie and Woodruff, Christopher (2013) 'The Demand for, and Consequences of, Formalization among Informal Firms in Sri Lanka' *American Economic Journal: Applied Economics* 5(2): 122-150

Fiszbein, Ariel and Norbert Schady (2009) *Conditional Cash Transfers: Reducing Present and Future Poverty* Washington: World Bank

Goertz, Gary and James Mahoney (2012) *A Tale of Two Cultures: Qualitative and Quantitative Research in the Social Sciences* Princeton, NJ: Princeton University Press

Greene, Jennifer C., Valerie J. Caracelli, and Wendy F. Graham (1989) 'Toward a Conceptual Framework for Mixed-Method Evaluation Designs' *Educational Evaluation and Policy Analysis* 11(3): 255-274

Henrich, Joseph, Steven J. Heine and Ara Norenzayan (2010) 'The Weirdest People in the World?' *Behavioral and Brain Sciences* 33: 61-135

Hentschel, Jesko (1999) 'Contextuality and Data Collection Methods: A Framework and Application to Heath Service Utilization' *Journal of Development Studies* 35(4): 64-94

Honore, Anthony (2010) 'Causation in the Law", in *Stanford Encyclopedia of Philosophy*. Available at: http://stanford.library.usyd.edu.au/entries/causation-law/

Hopkins, Daniel J. and Gary King (2010) 'Improving Anchoring Vignettes: Designing Surveys to Correct Interpersonal Incomparability' *Public Opinion Quarterly* 74(2): 201-222

Kauffman, Stuart A. (2016) *Humanity in a Creative University* New York: Oxford University Press

Kitzmuller, Markus and Martha Martinez Licetti (2012) 'Competition Policy: Encouraging Thriving Markets for Development' World Bank View Point Series. Available at http://siteresources.worldbank.org/EXTFINANCIALSECTOR/Resources/282884-1303327122200/VP331-Competition-Policy.pdf (Accessed 29 November 2016)

Kristjanson, Patti, Nelson Mango, Anirudh Krishna, Maren Radeny and Nancy Johnson (2010) 'Understanding Poverty Dynamics in Kenya." *Journal of International Development* 22(7): 978-996

Ludwig, Jens, Jeffrey R. Kling and Sendhil Mullainathan (2011) 'Mechanism Experiments and Policy Evaluations' *Journal of Economic Perspectives* 25(3): 17-38

Mansuri, Ghazala and Vijayendra Rao (2012) *Localizing Development: Does Participation Work?* Washington, DC: World Bank

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

Pawson, Ray (2006) *Evidence-based Policy: A Realist Perspective* London: Sage Publications

Pitt, Mark and Shahidur Khandker (1998) 'The Impact of Group-Based Credit Programs on Poor Households in Bangladesh: Does the Gender of Participants Matter?' *Journal of Political Economy* 106(5): 958-996

Pritchett, Lant (2013) 'The Folk and the Formula: Fact and Fiction in Development' Helsinki: WIDER Annual Lecture 16

Ragin, Charles (2014) *The Comparative Method: Moving Beyond Qualitative and Quantitative Strategies* (2nd edition) Berkeley, CA: University of California Press

Rao, Vijayendra (2002) 'Experiments in Participatory Econometrics: Improving the Connection Between Economic Analysis and the Real World' *Economic and Political Weekly* 22(20): 1887-1891

Rao, Vijayendra, Kripa Ananthpur and Kabur Malik (2017) 'The Anatomy of Failure: An Ethnography of a Randomized Trial to Deepen Democracy in Rural India' *World Development* 99(11): 481-97

Rao, Vijayendra and Michael Woolcock (2003) 'Integrating Qualitative and Quantitative Approaches in Program Evaluation', in Francois J. Bourguignon and Luiz Pereira da Silva (Eds.) *The Impact of Economic Policies on Poverty and Income Distribution: Evaluation Techniques and Tools* New York: Oxford University Press, pp. 165-90

Ravallion, Martin (2000) 'The Mystery of the Vanishing Benefits: An Introduction to Impact Evaluation' *World Bank Economic Review* 15(1): 115-140

Ravallion, Martin (2009) 'Evaluation in the Practice of Development' *World Bank Research Observer* 24(1): 29-53

Roodman, David, and Jonathan Morduch (2014) 'The Impact of Microcredit on the Poor in Bangladesh: Revisiting the Evidence' *Journal of Development Studies* 50(4): 583-604

Rugh, Jim, Michael Bamberger, and Linda Mabry (2011) *RealWorld Evaluation: Working Under Budget, Time, Data, and Political Constraints*. Thousand Oaks, CA: Sage Publications

Ruzzene, Attilia (2012) 'Drawing Lessons from Case Studies by Enhancing Comparability' *Philosophy of the Social Sciences* 42(1): 99-120

Sana, Mariano, Guy Stecklov and Alexander A. Weinreb (2012) 'Local or Outsider Interviewer? An Experimental Evaluation'. Available at http://paa2012.princeton.edu/papers/122313 (Accessed November 29, 2016)

Shaffer, Paul (2011) 'Against Excessive Rhetoric in Impact Assessment: Overstating the Case for Randomised Controlled Experiments' *Journal of Development Studies* 47(11): 1619-1635

Shanahan, Murray (2008) 'Dynamical Complexity in Small-World Networks of Spiking Neurons' *Physical Review E* 78(4): 041924

Teddlie Charles and Fen Yu (2007) 'Mixed Methods Sampling: A Typology with Examples' *Journal of Mixed Methods Research* 1(1): 77-100

Widner, Jennifer, Michael Woolcock and Daniel Nieto Ortega (eds.) (forthcoming) *The Case for Case Studies: Integrating Scholarship and Practice in International Development*. New York: Cambridge University Press

Woolcock, Michael (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, Michael (2013) 'Using Case Studies to Explore the External Validity of Complex Development Interventions' *Evaluation* 19(3): 229-248

Woolcock, Michael, Simon Szreter and Vijayendra Rao (2011) 'How and Why Does History Matter for Development Policy?' *Journal of Development Studies* 47(1): 70-96