Journal of Development Effectiveness
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t906200215

Toward a plurality of methods in project evaluation: a contextualised approach to understanding impact trajectories and efficacy

Michael Woolcock [a]
[a] University of Manchester, Brooks World Poverty Institute, Humanities Bridgeford Street Building, Manchester, UK

Online Publication Date: 01 March 2009

To cite this Article Woolcock, Michael(2009)'Toward a plurality of methods in project evaluation: a contextualised approach to understanding impact trajectories and efficacy',Journal of Development Effectiveness,1:1,1 — 14
To link to this Article: DOI: 10.1080/19439340902727719
URL: http://dx.doi.org/10.1080/19439340902727719

Journal of Development Effectiveness
Vol. 1, No. 1, March 2009, 1–14

Routledge
Taylor & Francis Group

Toward a plurality of methods in project evaluation: a contextualised approach to understanding impact trajectories and efficacy

Michael Woolcock*

University of Manchester, Brooks World Poverty Institute, Humanities Bridgeford Street Building, Oxford Road, Manchester M13 9PL, UK

Understanding the efficacy of development projects requires not only a plausible counterfactual but also an appropriate match between the shape of impact trajectory over time and the deployment of a corresponding array of research tools capable of empirically discerning such a trajectory. At present, however, the development community knows very little, other than by implicit assumption, about the expected shape of the impact trajectory from *any* given sector or project type, and as such is prone to routinely making attribution errors. Randomisation per se does not solve this problem. The sources and manifestations of these problems are considered, along with some constructive suggestions for responding to them.

Keywords: development planning and policy; economic change; impact assessment methods and approaches

1. Introduction

In most discussions among development researchers about how to enhance the efficacy of foreign aid and the quality of project evaluation, the emphasis is on how to upgrade the 'rigor' of the tools and techniques used to assess impact – that is, on methodological aspects of the evaluation protocol itself. Even in (rare) moments of critical self-reflection, advocates of putatively 'gold standard' techniques – that is, randomisation (Banerjee 2007) – couch their concerns primarily in terms of the ethics, logistics or political feasibility of assigning actual or potential project participants to 'treatment' and 'control' groups. At the end of the day, however, there remains the strong belief that gold standard protocols are the most (and for true believers, only) defensible basis on which to make hard decisions about whether a given project or policy is 'working' and thus whether or not it should be continued (terminated), scaled up (down) and/or replicated elsewhere.

In this paper, I seek to highlight the importance of striving for a better understanding of the project's known or likely impact trajectory over time. Specifically, I will argue that, in virtually all sectors, the development community has a weak (or at best implicit or assumed) understanding of the shape of the impact trajectories associated with its projects, and even less understanding of how these trajectories vary for different *kinds* of projects operating in different contexts, at different scales and with varying degrees of implementation effectiveness; more forcefully, I argue that the weakness of this knowledge greatly compromises our capacity to make accurate statements about project impacts, irrespective of whether they are inspired by 'demand' or 'supply' side imperatives, and even if they have been subject to the

*Email: michael.woolcock@manchester.ac.uk

ISSN 1943-9342 print/ISSN 1943-9407 online

DOI: 10.1080/19439340902727719
http://www.informaworld.com

most deftly implemented randomised trial. I contend that a truly rigorous evaluation is one that deploys the best available assessment tools at intervals that correspond to the shape of a project's known (via experience, empirical evidence, or inferred on the basis of sound theory) impact over time. For the purposes of simplicity of exposition, I shall call the shape of this net benefits profile over time the project's (or policy's) 'functional form' – that is, the impact trajectory that reflects the underlying 'technology' of the project and that is deemed to be (in effect) independent of scale, context and implementation effectiveness.[1]

The paper proceeds as follows. In Section 2, I spell out the nature and severity of this problem in more detail, locating it in the broader context of decision-making in development policy. Having done so, I then seek, in Section 3, to suggest some strategies for responding to it. Section 4 concludes.

2. Two big problems, one inadequate response

In contemporary development debates, and especially in discussions about project efficacy, the functional form for virtually all policies and projects – from rural roads and urban sanitation to guaranteed work programmes and microfinance initiatives – is effectively assumed to be (net of everything else) monotonically increasing and linear (that is, $y = mx + c$, as in the dashed line in Figure 1[2]), with the only serious methodological issue left to determine being the difference between the benefits obtained by the project participants and the counterfactual ('control') group, net of both groups' 'initial conditions'.[3] Put another way, when t_1 is actually calculated is largely irrelevant: with a functional form that is assumed to be monotonically increasing and linear, the net benefits to project/policy beneficiaries will be deemed to be 'significant' at the statistical moment when the difference between Y_1 and Y_1* passes the 0.05 level.[4] Such an assumption of impact trajectory, moreover, allows project managers and policymakers to extrapolate the flow of benefits (far) into the future. What matters most, then, is discerning the slope of the line – that is, correctly identifying 'c'. Obsessing about identification issues has been a feature of development microeconomics over the past decade, with randomisation given 'gold standard' status in large part because of its capacity to enhance the probability of correct identification.[5]

But it is only the most ad hoc theorising or wishful thinking (or the overriding imperatives of domestic political cycles and the structure of career paths at development organisations) that could possibly substantiate an assumption that *all* project impacts are linear and

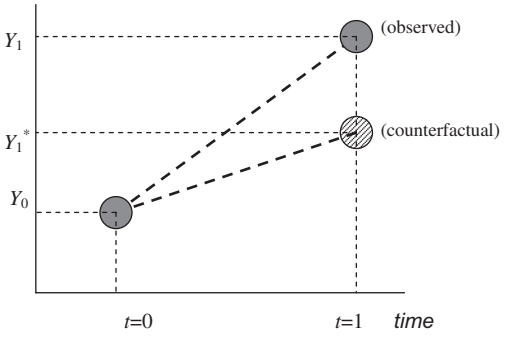

Figure 1. The canonical basis for assessing project/policy impact.

monotonic. For many (if not most) of the projects in its repertoire, the development community actually has a very weak substantive understanding, ex ante or even ex post, of what the functional form of those projects is, was, or might be.[6] We know we need 'baseline' (at time t_0) and follow-up data (at time t_1), but the content and shape of the proverbial 'black box' connecting these data points remains wholly a mystery, to the development industry's peril.

This issue is especially acute for 'social development' projects, such as those trying to increase the participation of marginalised groups, enhance women's empowerment or improve the accessibility of formal legal systems to the poor.[7] Even a cursory reading of social theory, for example, would suggest that in fact the most likely shape of such projects' functional form is a J-curve (that is, things get worse before they – hopefully, maybe – get better) or a step function (that is, long periods of stasis followed by a sudden rupture brought on by, say, an election or the reaching of a 'tipping point' in the adoption of a new fertiliser technology, in which prevailing norms and/or uptake by an influential local leader rapidly leads others to do likewise).[8] Moreover, one can reasonably imagine that, in reality, all manner of functional forms actually characterise the different types of available (or possible) development interventions: in addition to the above examples, anti-corruption efforts (as assessed, for example, via public expenditure tracking surveys) may have high initial impact that fades over time (as the bad guys figure out alternative ways to cream off funds); the trajectory of AIDS awareness campaigns may resemble an 'S' curve, in which there is slow initial uptake, a subsequent rapid increase as broader cultural norms and taboos are transgressed, and then a plateau as the final hard-to-reach populations are eventually contacted. (The appendix offers some speculative prognostications on the possible shape of the impact curve associated with various different types of development interventions.)

Development theorists, evaluators, project managers and senior administrators alike need to be explicit in articulating and substantive in defending the functional form that they believe characterises their interventions.[9] The absence of such knowledge can lead to potentially major Type I and II errors in assessing the efficacy of projects: for example, if a women's empowerment project does indeed have a J-curve functional form (men initially resist ceding resources and status, perhaps violently at first, only to come around when their attitudes and/or interests change, or prevailing local norms shift), and a 'gold standard' randomised evaluation happens to be conducted when the programme is at the bottom of that curve, it would be unceremoniously (but very inaccurately) deemed a failure. Similarly, efforts to enhance human rights may take many decades to be achieved – as did, for example, efforts to end slavery (Brown 2006) and judicial torture (Hunt 2007)[10] – and realising them may not only entail multiple setbacks but also (of course) vastly outlive the careers of even generations of leaders. In a professional environment that rewards rapid and easily measured development gains, however, managers of projects delivering such gains are likely to be promoted long before their equally diligent and talented colleagues slogging away faithfully on projects that are 'intrinsically' important and 'demand driven' (that is, reflecting a community's priorities) but inherently years away from being able to show demonstrable positive results.

Similarly, in the social psychology of the board room, where hard decisions have to be made about how to allocate finite development resources, directors are going to have a much easier time being persuaded that funds given to build roads, enhance irrigation and immunise children will produce positive, measurable and immediate impacts, certainly when competitors for these same funds are proposing to address land reform, consolidate peace accords, or initiate efforts to improving the judiciary in 'failed' states. For these latter projects, which

may be the country's highest priority, the metrics of success are inherently unclear and the functional form of the interventions to bring them about is largely unknown (perhaps even unknowable). By extension, projects with a known 'J' curve functional form (that is, where managers know that, for sure, things will get worse before they get better, such as hosting a truth commission to redress atrocities committed against indigenous groups) are hugely disadvantaged; for such projects, managers only have an incentive, ex ante, to disavow the likelihood of such a trajectory, preferring instead to argue that their projects too can generate positive, immediate, and measurable results.[11] To break this cycle likely requires a compelling, coherent, alternative moral vision of how organisational priorities will be determined, enacted and assessed (Heifetz 1994, Mintzberg 2004); from a researcher's perspective, however, it should also entail raising awareness of the reality that different kinds of projects are needed to respond to a range of different (idiosyncratic) country problems, and that the effective use of limited development resources will necessarily entail getting an appropriate match between problems, response options, impact trajectories, and a range of assessment tools capable of discerning them.

As things currently stand, however, short political attention spans, organisational imperatives to produce 'results', and international mandates to achieve 'targets' (such as the Millennium Development Goals) generate a net effect in which the development industry ends up reverse engineering itself, strongly preferring 'high initial impact' projects over projects that might actually respond to the problems that poor countries themselves deem a priority but which are inherently complex, hard to measure and/or necessarily slow to demonstrate positive impact. In short, the development profession strongly prefer to sell known, universal solutions with high, immediate and readily measurable impact rather than wrestle with ambiguous, context-specific problems that may not have (at least ex ante) a known or even knowable solution. The prevailing institutional and political imperatives largely predispose us to 'see' problems – to render them 'legible' (Scott 1998) – in terms of the models, assumptions and instrumentalities[12] we happen to have.[13]

It is important to stress that this approach will work for (indeed, is ideally suited to) some problems in some contexts (that is, those problems that are technical in nature, such as hyperinflation and road building, for which there is a technical solution that can be divined by experts and implemented by a coherent bureaucratic apparatus), but not all problems in all contexts. As such, the challenge for development theorists and practitioners alike is to craft mechanisms that (a) help forge a more appropriate match between *types* of problems and *types* of solutions, and (b) make any such pronouncements on the efficacy of these 'solutions' on the basis of some informed sense of the conditions under which certain levels of impact may be achieved *over time*. Here I am focusing on the second of these two issues,[14] but it should be clear that both are halves of a whole with respect to development effectiveness: the 'wrong' project offered in response to a given problem is unlikely to redress it (even if the project itself independently generates a 'positive' impact), and ignorance about the likely trajectory of impact associated with *any* project, over and above that of the counterfactual(s), renders suspect verdicts regarding its putative success, indifference or failure.

It should also be clear that randomisation, in and of itself, does not solve this core problem. Applied uncritically, it can even be said to enhance the problem, because of the surety with which some of its (uncritical) protagonists render 'definitive' judgements on projects on the basis of their 'rigorous', 'gold standard', 'scientific' evidence. Before going further, I should stress forcefully that I am certainly not 'against' randomisation; on the contrary, as a matter of principle and where a given project happens to lend itself to being randomised, I endorse politically feasible, administratively supportable and ethically acceptable efforts to use randomisation in project evaluation for all the standard reasons. I have

contributed to several such evaluations myself, and fully appreciate the efforts of those who, through championing randomisation, have raised the status, frequency and quality of project evaluation. For present purposes, my concern is not with randomisation per se, much less quantitative methods more generally, but rather with (a) claims made by anyone about the impact of a project that fail to take cognisance of where observed impacts reside on a longer impact trajectory curve (known or inferred), and (b) claims by champions of randomisation that do not acknowledge either the inherent limits of this method and/or that regard evidence derived from non-randomised protocols as inherently suspect ('anecdotal') and/or of second-order importance.[15]

It is my central contention that a truly rigorous evaluation is one that deploys the full arsenal of social sciences methods as part of a strategy focused on achieving – within the prevailing political, economic and logistical constraints – an optimal match between these methods (or combination of methods) and the *type* of problem to which the project (or policy) is responding. The policy problem must generate the methodological response, not the other way around, just as the available policy/project 'solutions' should not determine which problems are addressed. Put differently, individual methods per se are not 'rigorous'; randomisation is not *inherently* a generator of superior data. Rather, methods become rigorous when they comprehensively generate valid and reliable data that is able to speak to the specific characteristics of the problem in question. Randomisation is ideally suited to addressing one major form of bias, namely selection bias, and as crucial as this is, responding in a truly scientific way to the evaluation challenge in any given context requires responding to the many and varied sources of bias(es).[16]

The nature and salience of these biases become more apparent with projects adopting a 'community-driven' (participatory) approach, whose component elements violate many of the assumptions that make randomisation possible – for example, the unit homogeneity assumption, which holds that the 'treatment' being given is absolutely identical in terms of its content, size, frequency and duration of exposure for each recipient (Whiteside *et al.* 2006).[17] Participatory projects,[18] by design, adapt themselves to the idiosyncrasies of the contexts in which they operate; there are common principles that inform their design, and of course formal operational guideline documents that spell out what is to be done by whom, but considerable discretion is also given to front-line staff to interpret these documents on the basis of their professional judgement: they are instilled to take seriously the spirit of the project guidelines but to tailor them in response to the particular characteristics of the local circumstances in which they find themselves. In effect, 'the project' as experienced by participants is as varied as the range of contexts in which it is implemented.

The process of adapting the project to local contextual realities, however, clearly requires considerable skill and professionalism on the part of front-line staff, and this too will vary, especially since many of the staff are relatively young (usually in their mid-to-late twenties) – some will be especially diligent and gifted at working with communities, while others will not. Such staff are not functionaries dutifully providing a standardised service, such as immunising babies or distributing food rations; they are instead engaged in extensive face-to-face interaction with villagers over many months, making innumerable discretionary decisions. In many respects 'the project' is itself a dynamic decision-making process rather than a static 'product', and as such attempts to make causal claims regarding overall impact must address endemic unobserved heterogeneity bias. In short, on both the 'demand side' (local context) and the 'supply side' (front-line project staff) there is, *by design*, enormous variation, much of which cannot be readily observed through orthodox large-scale evaluation tools such as surveys.[19] Moreover, external validity claims – the imperative to believe that if a project works 'here' it can also work 'there' – are also made problematic in these

projects. In a targets-driven world, organisations will be strongly predisposed to favour development 'technologies' – that is, 'best practice' projects that can be safely assumed to deliver the same outcome independent of context and implementation effectiveness – because their impacts can be readily predicted, costed, extrapolated and managed; those projects that are inherently and especially dependent on context and implementation effectiveness, even if ultimately the most important from a recipient's perspective, will face a perennial uphill battle for resources, attention and legitimacy.

Another way in which social development projects present a challenge to evaluators and managers is grappling with the imperative to extrapolate not just from here to there, but from small to large. Some of the impetus for randomisation, for example, comes from the otherwise commendable adage to start small, rigorously assess, and then 'scale up' those interventions that are deemed initially successful (cf. Easterly 2006).[20] But will bigger be better? Those committed to achieving predetermined targets want to believe so, since it greatly simplifies the managerial challenge: just take 'what works' in the shortest time-frame and multiply it by whatever factor is needed to reach the goal. Again, this kind of assumption is relatively safe with standardised interventions such as malaria nets, but it is much less so for projects whose efficacy turns on context specificity and implementation effectiveness. In such instances, for example, the quality of staffing may be greatly influenced by scale: the proverbial 'best and brightest' may only join a project once it has reached a certain level of scale and prestige, while the entrepreneurs, innovators and risk-takers may only thrive in a relatively small operation. Similarly, a small 'empowerment' project may ruffle few political feathers when it is small and when senior managers need to show that they support a diverse portfolio of projects; when such projects rise to national prominence (as they have in Indonesia and Brazil), however, the political economy surrounding them – that is, the resources they command and political profile they attain – renders them rather different entities. Focusing solely on the results from a randomised evaluation of a pilot project as the arbiter for whether it should be 'scaled up' is thus problematic, in general but especially with participatory projects. Projects heavily dependent on the social skills of front-line staff are also likely to exhibit strong learning-by-doing effects that are enhanced by scale: initial forays by individual staff into village life may be awkward, generating only modest results, whereas subsequent experiential learning and collective sharing of 'lessons' may yield disproportionately higher impact.[21]

In the section that follows, I provide some concrete suggestions for responding to this critique of project evaluations generally, and of randomisation alone as the empirical arbiter for making decisions about the efficacy, scale and replication of development projects, especially those exhibiting a defining 'participatory' component. These suggestions stress the importance of crafting strategies that complement the strengths, and substitute for the weaknesses, of randomisation, using a broader portfolio of qualitative and quantitative methods.

3. What to do?

The logic of the analysis thus far suggests that a key evaluation challenge is identifying not only a counterfactual – that is, the outcomes that would have been obtained, all other things being equal, were participants not to have received the project – but generating a defensible sense of the underlying impact trajectory of the project intervention – that is, the outcomes that one would expect at a given point after the intervention began, given the type of project and the nature of the context in which it is operating. Identifying the counterfactual is an increasingly familiar staple of evaluation debates, but how might one undertake the task of discerning the shape of an impact trajectory?

There are at least three entry points, each of increasing degrees of sophistication. The first is simply raw experience: seasoned project managers should have a good sense of how long and in what form the impacts associated with a particular project in a particular context should take to materialise. It would be a step forward for development effectiveness, if not for research methodologists, if difficult decisions over the allocation of finite resources could incorporate a sensibility more overtly primed to the recognition that – assuming comparable levels of implementation quality – different types of projects are likely to yield different, sometimes initially negative, outcomes over different time periods. Even with no formal evidence to hand, seasoned project managers and their organisational directors should at least be able to draw on their career experience to help inform such deliberations.

Astute intuition and seasoned field experience combined with solid theory should provide a second avenue: the very essence of a good theory should be that it provides a sense and a justification of the conditions under which, and mechanisms by which, certain project outcomes should be expected. Child heath interventions, for example, are based on medical science; indeed, the very definition of a bio-medical intervention such as immunisation against polio is that the drugs in question work the same way in all humans everywhere. Success in one country or community is thus a valid and reliable basis on which to extrapolate to others, and public health officials can plausibly predict the impacts that a given level of funding will procure. More recently, interventions such as conditional cash transfers (especially in Latin America) have been subjected to all manner of randomised trials, but the implicit theory of impact trajectory on which they rest is one of monotonic linearity (see Appendix, graph A). This may well turn out to be empirically true, and in the absence of the necessary evidence it may also be eminently defensible theoretically, but a formal and explicit articulation of a theory of impact trajectory is needed nonetheless if a correct verdict on impact is to be rendered. Education projects arguably come closest to formally articulating a theory of impact trajectory, perhaps because it is clear to everyone that there is inherently a long lag between investments in schooling and the attainment of broader social outcomes (such as reduced infant mortality and higher incomes), though education is often also justified and promoted (correctly) on intrinsic grounds – that is, because it matters for its own sake, independently of whether it generates other ('instrumental') development outcomes.[22]

A defensible theory is most important, however, for participatory development projects, precisely because the nature of their impact trajectory is most likely to be different from other (more orthodox) projects, and indeed from each other. As argued above, the accumulated wisdom of social theory over two hundred years strongly posits that institutional change is wrought with conflict, and often proceeds fitfully: hard won gains may endure for years, only to be eroded or eliminated altogether. Punctuated equilibriums and 'J' curves are the likely shape of project trajectories seeking to empower marginalised groups, yet exactly how long the periods of stasis or how deep and enduring the bottom of the curve will be as varied as the idiosyncrasies of the context in which they take place. The theory will also need to take account of the fact that both the project and the context itself are likely to evolve over time – that is, strictly speaking, 'the project' unveiled at the start may not be the same 'project' that is present five years in, and the context itself may have changed by the very presence of the project (for example, by attempting to give women a stronger voice in local resource allocation decisions). Careful and extended engagement with each context, not hopeful extrapolation on the basis of results obtained elsewhere, should be basis on which such a local theory is crafted and, in turn, subsequent evaluation decisions are based.[23]

Thirdly, the regular collection of empirical evidence can itself be a basis for determining the shape of the project's impact trajectory, and is ultimately (for researchers at least) the

most defensible basis on which to do so. This can potentially be done as part of the project's monitoring procedures or (much less frequently) perhaps as part of an accompanying research exercise, but is likely to be extremely costly (and thus rarely done). The logistics and ethics of doing this as part of a randomised trial are also likely to militate against it, even if, in principle, it would actually constitute the 'diamond' (that is, higher than 'gold') standard for project evaluation.[24]

For researchers and/or advisors to those establishing project monitoring systems, it is important to generate a *range* of evidence, which in turn requires collecting, collating and interpreting qualitative and quantitative data (Rao and Woolcock 2003). The relative proportion of each required will depend on the nature of the project itself: the impact of roads and irrigation on local agricultural prices can be readily assessed with solid household (and other) survey data, while the effects of including more women in village meetings on the equity of resource allocation decisions and the quality of local dispute resolution systems requires more extensive engagement with qualitative data (Gibson and Woolcock 2008). Where it made sense – that is, where the nature of the project was (or component aspects of it were) such that it could be randomly assigned without 'cross over' effects between treatment and control groups – and was politically supportable, an ideal approach would be to incorporate mixed methods techniques as part of a randomised design, but as indicated above, something less than 'ideal' should not necessarily be regarded as 'less rigorous': such judgements should be rendered instead on the extent to which one has responded comprehensively to the problem at hand within the prevailing constraints. In any event, the challenge is to be as explicit, clear and substantive as possible about what the project's impact trajectory should look like, and to assess outcomes against this; to only compare against a counterfactual and a baseline measure is to risk making serious attribution errors.

A concrete example helps to show how a mixed methods randomised design can work in practice. For the last seven years, a team in Indonesia has been developing a series of small pilot projects designed to enhance the accessibility and quality of local legal services for the poor, especially poor women. Through a series of interventions designed to provide these women with identity registration documents (that is, paperwork that enables them to assert the most elementary aspects of citizenship, namely to prove who they are, when they were born, and so forth) and to have access to legal intermediaries able to help them move between the words of state law and customary law, the goal has been to enhance the likelihood that poor people will be able to assert their rights and obtain justice in everyday disputes (for example, over inheritance claims and land boundaries). The successes and failures of these pilots have been documented by extensive qualitative research, which have both enabled ongoing refinements to be made to the programmes while also clarifying the nature of the very problem that they are trying to rectify. Seven years in, the programme's managers have a solid sense of what works and what doesn't, how long it takes for favourable outcomes to emerge, are well versed in the idiosyncrasies of the contexts in which they operate, and have succeeded in raising the profile of their work to the point that senior legal figures in the country understand and support it, and are eager to expand it.

In 2006, the process of 'scaling up' was made possible by appending the programme onto a larger community development project operating in some of the poorest (and least stable) areas of the country. Working in partnership with local politicians and officials, it was determined that it would be possible and desirable to allocate this larger project on a randomised basis at the sub-district level. This process of randomised assignment will be accompanied by further qualitative research teams, who will explore in more detail the precise mechanisms by which local actors adopt, resist or work around these initiatives to try to improve the quality of justice for the poor. It remains to be seen whether and how both the project and the innovative

evaluation strategy plays out, but at least in principle it is an attempt to get the sequencing and content right: begin with a high-priority problem (that is, highly inequitable local justice and governance systems) discerned on the basis of intensive field research, design a context specific pilot project response, conduct accompanying evaluation research, articulate an informed sense of what the impact trajectory looks like and which factors most influence it, scale up at the appropriate time, and then use the best possible evaluation procedure (mixed methods randomised design) to assess larger impacts.[25]

4. Conclusion

In the context of widespread popular ignorance about the scale of foreign aid[26] and enduring scepticism regarding the extent to which it yields positive and 'sustainable' impacts, debates between those advocating local experimentation (Easterly 2006) over grand plans (Sachs 2005) serve a useful purpose by raising awareness, clarifying the nature and basis of diverse perspectives, and encouraging more hard-nosed efforts to understand 'what works and why' (World Bank 1998, Riddell 2007). A remarkably small percentage of development initiatives have actually been formally evaluated,[27] but even amongst those committed to bringing a more comprehensive evidence base to bear on these debates, the terms of debate have tended to coalesce around the merits of various econometric strategies for resolving particular identification issues, on the one hand, with token genuflection to the importance of 'context', on the other.

This is, in its own way, an importance advance, but in the process it commonly makes assumptions about the nature of impact trajectories over time, namely, that they are monotonically linear. This assumption, I have argued, is not only unsubstantiated empirically in most cases, but at best is likely to apply to only a relatively narrow class of interventions. Experience, theory (especially social theory) and evidence itself, however, suggests that impact trajectories are likely to vary considerably across different *types* of development intervention, with corresponding consequences for the veracity of claims made pertaining to impact assessment. This challenge is likely to be especially salient for interventions (for example, participatory development projects) that, explicitly and by design, seek to adapt themselves onto the idiosyncrasies of local context – that is, where the intervention is deliberately non-standardised, where 'the project' is, in effect, many projects. The problems are further compounded, moreover, to the extent the presumed efficacy of such interventions turns on their implementation effectiveness. At its core, internal and especially external validity claims suffer greatly from the absence of knowledge not just of counterfactuals, but of where a given performance outcome at a particular point in time is located vis-à-vis where it should (or might) be expected to be.

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 and when this variance manifests itself is crucial for making accurate empirical evaluations of project/policy effectiveness. As such, assessing the recipients of aid assistance according to whether they have achieved their stated 'goals' – and using such assessments as the basis for determining further streams of funding – can only be meaningfully undertaken, I contend, when it is made in the context of an explicit articulation of a given sector's and/or particular project's impact trajectory. Absent such knowledge, the capacity for fundamentally inaccurate conclusions to be drawn (or, more formally, for Type I and Type II errors to be made) is rife.

As such, performance-based project initiatives should be part of, rather than a substitute for, an approach to reformulating development strategies, a reformulation that stresses not only true 'country ownership' and a correct alignment of incentives for all actors involved, but also a more coherent sequencing of deliberations from identification of problems to exploration of possible solutions, to knowledge of how those solutions are likely to change as a function of time, scale, context and implementation effectiveness. Acquiring such knowledge will not be a product of simply deploying what some deem to be 'gold standard' evaluation protocols (for example, randomised trials) 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 historical) available to social scientists.

Acknowledgements

The paper reflects the views of the author alone and should not be attributed to the University of Manchester or the World Bank (from which the author is currently on external service leave). The author is grateful to the Center for Global Development for opportunities to discuss these issues in a specific policy context and to seminar participants at AusAID, Columbia University, WIDER and the World Bank for helpful comments. He also benefited immensely from frank discussions on these issues with Michael Clemens, Pascaline Dupas, Indranil Dutta, James Foster, Elizabeth Levy Paluck, Lant Pritchett, Vijayendra Rao, Caroline Sage, Sandra Sequeira and Martin Ravallion (though none should be implicated where there are remaining errors of fact or interpretation). Thanks also to the Journal's editor and two anonymous referees for helpful suggestions.

Notes

1. As I argue below, it is a very strong assumption that a development project does actually embody an invariant 'technology' (that is, more of X will generate less of Y over time period Z, independently of context, scale and implementation effectiveness); many participatory development projects, for example, are, explicitly and by design, *not* 'technologies' in this sense.
2. There are some important exceptions (for example, Moffitt 2006), but it remains the case that the dominant assumption in the vast majority of discussions about policy, project, and programme effectiveness assumes a linear functional form. Leading evaluation textbooks such as Weiss (1998) and Rossi *et al.* (2003) hint at, but do not explicitly address, the possibility that impact trajectories might be highly variable.
3. This is the classic 'double difference' protocol, but its logic informs the basis of all evaluation designs.
4. I leave aside here – though the thrust of my remarks on this point are surely consistent with – McCloskey's lifelong quest to implore economists (and by extension other social scientists) to appreciate the difference between statistical and economic significance (see, only most recently, Ziliak and McCloskey 2008).
5. On the general problem of identification in the social sciences, see (among others) Manski (1995).
6. An important paper by King and Berhman (forthcoming) reviews the evidence pertaining to the impact of health and education projects over time, but stops short of spelling out the fuller implications, for all development sectors, of having inadequate knowledge of such matters. See also Mu and van de Walle (2007) which deftly explores the heterogeneity of impacts across time and space associated with rural roads in Vietnam. The paper is important for its discernment of positive impacts some time after an initial assessment had found minimal evidence to this effect. Researchers are generally familiar with the likelihood of spatial and demographic heterogeneity of impacts (for example, Galasso and Ravallion 2005); it is much less frequent that the possibility of considerable variation is considered over time.
7. On the broader empirical challenges of assessing participatory projects, see Mansuri and Rao (2004).

8. Related arguments are made in the field of evolutionary biology, for example, where Gould (2007) famously (if controversially) argues forcefully for the notion of 'punctuated equilibrium' – species stay the same for long periods of time, then change rapidly in response to an external shock or a mutant gene (for example, changes in pigmentation) that proves qualitatively superior in addressing a particular problem (avoiding predators). See also Thomas Kuhn's (1968) (even more famous and controversial) arguments that 'paradigm shifts' in science (and elsewhere) follow a similar step-function trajectory, with orthodoxy entrenching itself and then prevailing well beyond its time as an old guard initially succeeds in defending it but then dies off, with a new paradigm then rapidly taking its place.
9. It should be clear that this line of argument is different from the perennial practitioner's lament that development impacts 'take time'; some impacts may be immediate and others may indeed only be apparent after many years, but I am arguing here that it matters both empirically and politically what the *trajectory* of impacts associated with a given project takes over time.
10. Indeed, as Hunt (2007) carefully shows, the gradual realisation of human rights took centuries, not decades, and their expansion generated a corresponding powerful backlash (for example, eugenics in the nineteenth century). Human rights today, of course, are still far from being universal; an estimated seven million people remain in literal slavery, long after every country has officially condemned the practice.
11. Predictably enough, those engaged in judicial reform efforts around the world are now being challenged to come up with a range of indices for 'measuring justice'. There is nothing sinister in this, and as a researcher I am certainly, in most respects, all for more and better data; the core problem, rather, is that the relentless focus on 'measurement' for the sake of it masks the recognition that certain kinds of development policy problems (such as legal/judicial reform) require qualitatively different kinds of decision-making and response mechanisms to those in civil engineering and accounting, with correspondingly different ways of, and time horizons for, assessing success (failure). On the general problem of how prevailing organisational incentives mitigate against evaluation of *all* development projects, see Pritchett (2002).
12. Or sense-making apparatus, as some organisational theorists (for example, Weick 1995) would call it.
13. These sensibilities are on strong display in Sachs (2005).
14. Pritchett and Woolcock (2004), building on Scott (1998), explores analytical and operational aspects of the first problem in more detail.
15. For a related critique, see Ravallion (2008), who rightly stresses that our primary role as development researchers is to be useful within the constraints we find ourselves in, not limiting our research efforts to those relatively narrow class of projects that either can be or have been randomised.
16. Bio-medical research, for example, from which champions of randomisation in development claim to draw their inspiration, routinely deploys a triple blind placebo controlled protocol to address an additional four potential sources of bias. In such studies, neither the participants, those implementing the 'treatment' nor their supervisors knows who is in the control or treatment group and whether they have received the actual treatment or a placebo. These sources of bias are also salient for development, though one struggles to imagine how exactly one could actually implement a triple-blind placebo-controlled development project. Precisely because this is logistically impossible even as the multiple sources of bias endure, a *range* of data and methods needs to be called upon if the evaluation is (and the conclusions to which it gives rise are) to be deemed truly rigorous.
17. This does not preclude, of course, individual component elements of such programmes from being adjusted in randomised 'treatment' and control groups (see Olken 2007, Paluck 2008); it just makes it highly problematic to issue summary declarations on the overall impact of the project, since there really isn't 'a project', but rather tens of thousands of manifestations of a project that has been, by design, allowed to adapt itself to the idiosyncrasies of the contexts in which it is located.
18. For example, the Kecamatan Development Project in Indonesia (see Guggenheim 2006). See Mansuri and Rao (2004) for a thorough review of the characteristics and claims of such projects.
19. As is well known, unobserved heterogeneity bias is also endemic in 'matching' designs, such as those using propensity scores, which rely heavily on observed data.
20. I leave aside here the fact that it is hard to name a single successful major policy initiative in developed countries that came about as a result of this sequence; most were 'born large' or were

scaled up because initial efforts were deemed a political and/or administrative success, not an empirical one.

21. This has been the experience, for example, of the Kecamatan Development Project (KDP), now a nation-wide social development project in Indonesia and one of the largest of its kind in the world. Frequently declared a success on numerous grounds – for example, its capacity to reduce corruption in the building of rural roads (Olken 2007) – its initial performance in the pilot stage was modest at best; that is, it became a success *after* it reached scale. The initial decision to scale-up was in no small part a political one, but it was also the case that, having attained prominence, KDP attracted Indonesia's best young graduates, its most senior political managers, and was able to learn from its initial experiences. On the origins and structure of KDP, see Guggenheim (2006).
22. See also Clemens (2004), who usefully shows that, historically, aggregate education enrolments seem to display a consistent 'S' pattern – that is, low initial take-up, followed by a rapid expansion, then slow progress to full enrolment. Such knowledge adds another layer to the arguments advanced in this paper regarding project-level trajectories (and is, in turn, essential to understanding the prospects of attaining even higher-order education targets such as the second Millennium Development Goal).
23. This route into understanding impact trajectories would also strongly support closer engagement with historical methods of analysis, as complements to more orthodox qualitative and quantitative approaches. On the importance of history for development policy, see Woolcock *et al.* (2008).
24. The true 'diamond standard' evaluation, of course, would also incorporate a triple-blind placebo controlled experimental design, but the very impossibility of this for participatory projects warrants no further discussion.
25. Other examples of mixed methods in development research and project evaluation include Hentschel (1999), Bamberger (2000), White (2002), Rao and Ibanez (2005) and Jha *et al.* (2007). More generally, see the Q-squared project led by Ravi Kanbur and Paul Shaffer: http://www.q-squared.ca/.
26. Pollsters repeatedly find that citizens of OECD countries vastly overestimate the amount of money (as a percentage of the federal budget) given by their government for the purpose of foreign aid.
27. Then chief economist of the World Bank, Francois Bourguignon, was quoted in *The New York Times* in 2005 as saying that only '5 per cent' of World Bank projects had been subjected to fully rigorous evaluation.

References

Bamberger, M., 2000. *Integrating qualitative and quantitative research in development projects*. Washington, DC: World Bank.

Banerjee, A., 2007. *Making aid work*. Cambridge, MA: MIT Press.

Brown, C.L., 2006. *Moral capital: foundations of British abolitionism*. Chapel Hill: University of North Carolina Press.

Clemens, M., 2004. The long walk to school: international education goals in historical perspective. Working paper 37. Center for Global Development.

Dichter, T., 2003. *Despite good intentions: why development assistance to the Third World has failed*. Amhert, MA: University of Massachusetts Press.

Easterly, W., 2006. *The white man's burden: why the West's efforts to aid the rest have done so much ill and so little good*. New York: Penguin.

Galasso, E. and Ravallion, M., 2005. Decentralized targeting of an antipoverty program. *Journal of public economics*, 89 (4), 705–727.

Gibson, C. and Woolcock, M., 2008. Empowerment, deliberative development and local level politics in Indonesia: participatory projects as a source of countervailing power. *Studies in comparative international development*, 43 (2), 151–180.

Gould, S.J., 2007. *Punctuated equilibrium*. Cambridge, MA: Harvard University Press.

Guggenheim, S., 2006. Crises and contradictions: explaining a community development project in Indonesia. *In*: A. Bebbington *et al.*, eds. *The search for empowerment: social capital as idea and practice at the World Bank*. Bloomfield, CT: Kumarian Press, 111–144.

Heifetz, R., 1994. *Leadership without easy answers*. Cambridge, MA: Harvard University Press.

Hentschel, J., 1999. Contextuality and data collection methods: a framework and application to health service utilization. *Journal of development studies*, 35 (4), 64–94.
Hunt, L., 2007. *Inventing human rights: a history.* New York: Norton.
Jha, S., Rao, V. and Woolcock, M., 2007. Governance in the gullies: democratic responsiveness and community leadership in Delhi's slums. *World development*, 35 (2), 230–46.
King, E. and Behrman, J., forthcoming. Timing and duration of exposure in evaluation of social programs. *World Bank research observer.*
Kuhn, T., 1968. *The structure of scientific revolutions*. Chicago: University of Chicago Press.
Manski, G., 1995. *Identification problems in the social sciences*. Cambridge, MA: Harvard University Press.
Mansuri, G. and Rao, V., 2004. Community-based and -driven development: a critical review. *World Bank research observer*, 19 (1), 1–39.
Mintzberg, H., 2004. *Managers, not MBAs: a hard look at the soft practice of managing and management development*. San Francisco: Berrett-Koehler Publishers.
Moffitt, R., 2006. Forecasting the effects of scaling up social programs: an economics perspective. *In*: B. Schneider and S.K. McDonald, eds. *Scale-up in education: ideas in principle*. Lanham, MD: Rowman and Littlefield, 173–186.
Mu, R. and van de Walle, D., 2007. Rural road and poor area development in Vietnam. Policy research working paper 4340. Washington, DC: World Bank.
Olken, B., 2007. Monitoring corruption: evidence from a field experiment in Indonesia. *Journal of political economy*, 115 (2), 200–249.
Paluck, E.L., 2008. The promising integration of qualitative methods and field experiments. Mimeo. Cambridge, MA: Harvard University.
Pritchett, L., 2002. It pays to be ignorant: a simple political economy of rigorous program evaluation. *Policy reform*, 5 (4), 251–269.
Pritchett, L. and Woolcock, M., 2004. Solutions when the solution is the problem: arraying the disarray in development. *World development*, 32 (2), 191–212.
Rao, V. and Ibáñez, A.M., 2005. The social impact of social funds in Jamaica: a mixed-methods analysis of participation, targeting and collective action in community driven development. *Journal of development studies*, 41 (5), 788–838.
Rao, V. and Woolcock, M., 2003. Integrating qualitative and quantitative approaches in program evaluation. *In*: F.J. Bourguignon and L.P. da Silva, eds. *The impact of economic policies on poverty and income distribution: evaluation techniques and tools*. New York: Oxford University Press.
Ravallion, M., 2008. Should the randomistas rule? Mimeo. World Bank.
Riddell, R., 2007. *Does foreign aid really work?* New York: Oxford University Press.
Rossi, P.H., Lipsey, M.W. and Freeman, H.E., 2003. *Evaluation: a systematic approach*. 7th ed. Thousand Oaks, CA: Sage Publications.
Sachs, J., 2005. *The end of poverty.* New York: Penguin.
Scott, J., 1998. *Seeing like a state: how well-intentioned efforts to improve the human condition have failed*. New Haven: Yale University Press.
Weick, K., 1995. *Sensemaking in organizations*. London: Sage.
Weiss, C., 1998. *Evaluation*. 2nd ed. Upper Saddle River, NJ: Prentice Hall.
White, H., 2002. Combining quantitative and qualitative approaches in poverty analysis. *World development*, 30 (3), 511–22.
Whiteside, K., Briggs, X. and Woolcock, M., 2006. Evaluating community-driven development: integrating science and local action. Mimeo. Washington, DC: World Bank, Development Research Group.
Woolcock, M., Szreter, S. and Rao, V., 2008. How and why does history matter for development policy? Working paper 70. Brooks World Poverty Institute, University of Manchester.
World Bank, 1998. *Assessing aid: what works, and why.* New York: Oxford University Press.
World Bank, 2003. *World development 2004: making services work for poor people*. New York: Oxford University Press.
Ziliak, S. and McCloskey, D., 2008. *The cult of statistical significance: how the standard error cost us jobs, justice and lives*. Ann Arbor: University of Michigan Press.

Appendix. Possible impact trajectories for various development projects