The craft of evaluative practice: negotiating legitimate methodologies within complex interventions

Steve Connelly
University of Sheffield, UK
Dave Vanderhoven
Independent researcher, Sheffield, UK

Corresponding author:
Steve Connelly, Department of Urban Studies & Planning, University of Sheffield, Winter St., Sheffield S10 2TN, UK.
Email: s.connelly@sheffield.ac.uk

Abstract

Evaluations of complex interventions are likely to encounter tensions between different methodological principles, and between the inherent causal rationality of evaluation and the messy complexity of real institutional contexts. Conceptualising evaluation as producing putatively authoritative evidence, we show how ‘legitimacy’ is a useful concept for unpacking evaluation design in practice. A case study of service integration shows how different approaches may have unpredictable levels of legitimacy, based in contrasting assessments of their methodological acceptability and actual utility. Through showing how practitioners resolved the tensions, we suggest that crafting a patchwork of different methodologies may be legitimate and effective, and can be seen as underpinned by its own pragmatic rationality. However, we also conclude that the explanatory power of theory-driven evaluation can be embedded in such an approach, both in elements of the patchwork and as an overarching guiding principle for the crafting process.

Keywords

Legitimacy, theory-driven evaluation, evaluation practice, pragmatism, service integration

The craft of evaluative practice: negotiating legitimate methodologies within complex interventions

We've struggled with this evaluation; I'm not sure why we've struggled so much.

Lyndon project sponsor

Introduction

Evaluation design rests on a plethora of issues: the evaluation's purposes, its audience, the resources available and so on. These sit alongside often deep-seated beliefs in the value of different methodologies, each with their perceived strengths and weaknesses, protagonists and critics. In the context of evaluating complex social policy interventions (Sanderson 2000; Ling 2012), methodological choices are likely to be complicated and contested. In this paper our aim is to enhance understanding of how such choices are made, through what we conceptualise as dynamic meaning-making and learning processes, driven by multiple rationalities and with uncertain outcomes (Weiss 1979). Central to these is the 'interplay' between the technical aspects of evaluation design and the institutional and political nature of an evaluation's context (Saunders 2012).

Within a broadly neo-institutionalist approach (Sanderson 2000) we propose an innovative analysis which focuses on legitimacy: that is, whether an evaluation methodology can be taken as authoritative (Schmitter 2001). Despite legitimacy's universal role in stabilising social processes (Zelditch 2001), the legitimacy of evaluation and evaluation-derived evidence is curiously absent as an explicit concern in the literature. Here we operationalise the concept through a constructivist framework, based on Beetham's approach to the legitimation of political power (Beetham 1991). This takes us beyond simply recognising complexity, to provide analytical tools for dissecting negotiations over rival methodologies as processes of legitimation and de-legitimation, in which technocratic and political arguments are equally present.

Our approach is motivated and informed by close observation of an adult social care integration programme and the delivery team's struggle not just to evaluate it, but to work out how best to evaluate it. In common with many projects in this and other policy fields, the Lyndon Project[1] involved multiple interventions in complex processes, driven by the assumptions that organisational change leads to better outcomes and (in the UK at least) reduced costs. Given the importance of these goals, developing appropriate evaluation methodologies has attracted much attention, especially as the evidence supporting these assumptions is very weak (Wistow and Dickinson 2012; Baxter *et al.* 2017).

In the narrative presented below, we demonstrate how those involved successfully negotiated tensions over the legitimacy of different evaluation methodologies. We show that there were surprising and important differences between how people believed they

[1] 'Lyndon' is a pseudonym.

ought to evaluate and what they believed was most effective, and in particular between rhetorical commitments to quantitative outcomes measures and actual practices of qualitative process understanding. Given that we share with many academics a 'fascination' with theory-driven evaluation (TDE) (Sullivan and Stewart 2006: 195) we were particularly interested in whether we could use our advisory role in the programme to base the evaluation in a coherent theory of change (ToC). We more-or-less failed, and the project's final evaluation report was a patchwork of different kinds of evidence, which nevertheless commanded legitimacy.

We conclude that understanding the constructed and contested legitimacy of evaluation methodologies in practice is useful, not only for explaining outcomes but also in identifying possible approaches to effecting change. Evaluations should be seen as complex interventions in themselves, and therefore one cannot expect single methodologies to be taken up unproblematically: in practice, different rationalities may be combined. This suggests an appreciation of evaluation as a craft (Sanderson 2000), with its own pragmatic rationality which privileges situated judgements about legitimacy and utility over methodological principle. However, we argue that this still allows a significant role for the causal rationality of theory-driven approaches.

Before presenting the story of the case, we consider further how issues of complexity and interpretation create challenges for evaluation, and set out our development of Beetham's approach to analysing legitimacy.

Rationalities of evaluation

Evaluation practice has long been dominated by the quantitative assessment of outcomes, with the attribution of impact assessed as far as possible through designs inspired by the natural (and particularly biomedical) sciences - randomised controlled trials are the aspirational 'gold standard' (Lehmann 2015). These methodologies are rooted in a positivist 'modernist' rationality (Sanderson 2000) which links: an understanding of policy making and bureaucracies as inherently rational (Adelman 1996); assumptions of linearity both in the processes being evaluated (Wagenaar and Cook 2003) and in the translation of evaluation 'findings' into policy (Nutley *et al.* 2007); and a view of causality as (only) observable as the 'constant conjunction' of events (Pawson and Tilley 1997). However, such approaches have struggled to provide useful accounts of the effectiveness of policy interventions in tackling complex social problems, producing poor explanations which neglect the complexity of the institutional context both of interventions and of evaluations (Sanderson 2000). Moreover, the problem is self-reinforcing, as the 'everyday positivism' of policy makers (Wagenaar and Cook 2003) sustains the predilection for modernist evaluation, despite its manifest failings.

For many years (going back at least to Weiss and Rein in 1970) a section of the evaluation community has responded with methodologies which start from the principle that evaluation should go beyond assessing outcomes to provide causal accounts of the processes which explain how and why policy interventions work (or not) (Coryn *et al.* 2011). Modernist rationality has also been challenged by recent academic work drawing on neo-institutionalist approaches, which treat the rules and norms of organisational contexts as central to explaining the general failure of mainstream evaluation to create useful knowledge (Sullivan 2011; Saunders 2012; Højlund 2014). So, for instance, Saunders emphasises how the actual *use* (or not) of an evaluation is intimately connected with its *usability* – the extent to which its design 'maximizes, facilitates or disables its potential use' (Saunders 2012: 422) in a particular setting. These two strands converge in support of 'theory-driven evaluation' (TDE) of various kinds (Chen 1990; Coryn *et al.* 2011), with the claim that TDE's causal accounts are a better basis for design for usability, as they can support improved policies and programmes though organisational learning and increased accountability (Berriet-Solliec *et al.* 2014).

There may, however, be a fundamental conceptual problem which undermines the utility of evaluation, common to modernist outcomes evaluation and TDE. Evaluation is *essentially* rationalist: the concept assumes that a goal-oriented intervention of some kind can be identified, and that there is a causal connection between intervention and observable change (Maxwell 2012). These assumptions may fail in the messy real world of policy and practice: in complex, dynamic, open systems causally relevant beginnings and effects of interventions may not be identifiable (Sanderson 2000; Ling 2012). Moreover, in a context of multiple, potentially competing interpretations it may not be possible to create a singular theory of change (ToC), a potentially fatal problem for any theory-based solution to the attribution problem (Sullivan and Stewart 2006). The strongest response is to abandon the possibility of defining and explaining outcomes altogether (Sullivan 2011) – clearly problematic in a policy world which demands such knowledge. But this seems unnecessary, and here we follow Sanderson in the search for '*terra firma* between…the illusion of certainty in modernist-rationalist order and…the danger of a pessimistic nihilism when facing chaos and complexity' (2000: 445), though to a rather different destination.

We distinguish two aspects of the 'modernist-rationalist' paradigm: its understanding of policy making and bureaucracies as rational processes and organisations (Adelman 1996) and the more fundamental issue of evaluation being inherently causal. Regarding the first, we need to recognise the differences within the broad family of TDE, which encompasses everything from positivist, quantitative causal modelling (Solmeyer and Constance 2015) through to interpretive approaches which see meaning-making as central to policy making (Pouliot 2014). The former are clearly vulnerable to the anti-modernist argument, but the latter perhaps are not, if their understanding of causality (needed to sustain evaluation as a concept) is compatible with their interpretivist epistemology.

Here evaluation based philosophically in critical realism (CR) (Pawson and Tilley 1997) offers a way forward. CR conceptualises causality as based in inherent powers of agents and of the ideational and material structures which constitute the world. This intrinsically qualitative understanding stands in contrast to the positivists' 'successionist' view of causality as that which links and explains (quantifiable) regularities between events (Maxwell 2012: 35). CR tends to be epistemologically interpretivist, at least insofar as human agency is concerned. In theorising a complex policy process, this allows differing interpretations and rationalities to be treated as causal factors in their own right, rather than as conflicting positions which must be (but perhaps cannot be) reconciled in order to create a singular theoretical understanding of what is going on. We need, though, to be humble in our claims to have theorised a process (Sanderson 2000; Patton 2012), and strive not for full understanding but for a level of 'practical adequacy' (Sayer 1992) good enough for tentative explanation and modest prediction, and always open to revision[2].

However, causal theory-based approaches face more mundane challenges. TDE seems particularly resource-demanding, with extended engagement required for the intimate organisational knowledge and stakeholder involvement necessary to develop causal theories (Ling 2012; Ivaldi *et al.* 2015). It may thus not be practically feasible, particularly in the context of financial austerity (Sullivan and Stewart 2006; Montague and Porteous 2013). Further, while Ling argues that policy makers 'should not be too disappointed when [an evaluator's] answer begins with the words "it depends"' (Ling 2012: 82) real world policy makers need usable, generalisable knowledge about the relationship between interventions and outcomes (Sullivan 2011).

Sanderson's own destination in the search for *terra firma* is to treat evaluation as a craft, involving 'a range of methods *appropriate* to particular circumstances' (2000: 450, emphasis added). Similarly Patton uses the analogy of judicial process to encourage an evaluator to use 'conflicting and confused evidence, and [sort] it out as best they can to reach an informed and hopefully fair judgment based on the cumulative evidence' (2012: 267). The suggestion appears to be that successful practice might involve pragmatic selection of methods without necessarily having any underlying single rationality beyond judgements of 'appropriateness'. What is not clear, however, is the basis for such selection: what guides evaluators when they make a 'carefully balanced judgement in "constructing the intelligible picture"' (Sanderson 2000: 449)?

Legitimacy

To address this we need a way of examining how and why different rationalities are manifested in practice. Here we turn to the concept of legitimacy, which is fundamentally an issue of whether or not a social practice fits into a system of norms: that is, a socially shared framework defining what is right and appropriate (Zelditch 2001)[3]. In the context of the legitimacy of power, Beetham (1991) argues that such judgements involve more than simply consent based on a belief in legitimacy (as Weber (1968 [1922] influentially claimed).

[2] Sanderson (2009) suggestively links policy making in general (rather than evaluation) to American pragmatist philosophy. In terms of implications for method, the overlaps between critical realism's 'practical adequacy' and Dewey's work seem clear, as is their rejection of correspondence theories of truth. Whether other underlying ontological and epistemological commitments are compatible seems doubtful, but is an unfinished debate beyond the scope of this paper.

[3] Social practices are conceptualised not simply as actions, but as meaningful actions which are therefore intrinsically social and normative: recognition as meaningful involves categorisation and judgment (Barnes 2001; Schatzki 2001).

Instead, he argues that the legitimacy of regimes rests on the exercise of power in conformity with a particular system of norms: a set of ‘established rules’; ‘shared beliefs’ which justify those rules; and the ‘express consent’ of those subordinate to the regime in question (Beetham 1991: 19). His approach has the advantages that it allows for explanatory accounts of contestation and change in the legitimacy of relations of power, and provides an analytical framework (the triads of rules, justifying beliefs and consent) which render the abstract concept of ‘legitimacy’ into something observable and researchable.

While Beetham’s analysis has been criticised within political science as being simplistic in relation to the legitimation of regimes (O'Kane 1993), as a constructivist and institutionalist approach it seems very open to being developed for the analysis of multiple and competing legitimacy claims in complex contexts. Thus Connelly (2011) and Lau (2013) have shown how Beetham’s approach can be used in the context of local governance, to explore the issues of how different actors and policy processes are accepted and become influential. Here we suggest that the analytical framework can be pushed beyond political science’s concern with the legitimacy of actors and their exercise of power. By moving away from the Weberian idea that legitimacy is essentially and only about consent, Beetham introduces justification for rules on principled or abstract grounds (O'Kane 1993). This opens the way to use the framework to analyse other contexts in which judgements of the acceptability of power are made which draw on general principles, including those about the status of different forms of knowledge and the methodologies used to generate them. This would seem eminently suitable for an examination of evaluation practices, as these are - overtly at least – a way of creating authoritative knowledge about a process in order to guide further action.

‘Overtly’ is important: evaluations can serve purposes and functions beyond this instrumental use (Saunders 2012; Højlund 2014). Weiss (1979) discusses ‘political’, often unspoken, uses of evidence in the policy process such as neutralising opponents and deflecting criticism. In some cases evaluating, and being seen to evaluate, may be the most important, symbolic purpose (Saunders 2012), rather than the production of knowledge. However, the instrumental use is part of the core meaning of the concept of evaluation, in a way these other functions are not.

Applying Beetham’s framework means that if an evaluation methodology is to become legitimate it will need to: follow accepted rules, which are justifiable according to sufficiently widely shared beliefs, and command sufficient consent by those who are involved in the process. ‘Justifying beliefs’ is the most complex of these criteria, encompassing both process and outcomes. Legitimacy will rest on a mixture of people’s (often taken-for-granted) beliefs about what kinds of processes generate valid knowledge, and the kinds of data they think they need to achieve their goals for the evaluation. So while we might expect, for instance, randomised trials to command legitimacy in an ‘everyday positivist’ policy setting because of assumptions about ‘good science’, an academic evaluator in the same setting might attempt to follow the rules for producing a ToC, justified by the belief that causal models are essential for generating useful knowledge.

But legitimacy is not just a matter of the in-principle justifiability of design: it is also a pragmatic matter of whether a methodology is acceptable (i.e. consented to by those involved) because it produces practically usable knowledge (cf. Saunders 2012). Further, the issue of whether beliefs are 'widely-shared' enough to justify a choice is made more complex by the importance of assumptions about others' beliefs in creating and sustaining legitimacy (Johnson *et al.* 2006): judgements about usability are likely to depend on whether an evaluator thinks their audience will accept a proposed methodology.

Beetham makes it clear that legitimacy requires stability across the three 'dimensions' of rules, justifiability and consent, but that where alternative rules or justifying principles exist it is always open to contest. While his concern is principally with challenges to a regime, here we are interested in how competing methodologies fare. The previous section suggested that while modernist rationality may be dominant, there is no single evaluation methodology which commands universal consent, particularly given the many functions which evaluation can fulfil and the likelihood that different stakeholders will have different goals and 'frames of reference' (Saunders 2012: 430). In the context of evaluating complex interventions the possibility for disagreement over design and implementation is therefore likely to be high.

We seek to explain why an evaluation is carried out in a particular way – why one methodology is chosen rather than another - through examining the triads of governing rules, justifying beliefs and levels of consent which different approaches command. None of this can be predicted *a priori*. Such examination means paying attention to actual practices (Sullivan 2011), and an open mind about what constitutes a legitimating rule or belief. We thus avoid normative assumptions about what is legitimate (Hurrelmann *et al.* 2007): in particular, we do not assume that legitimacy derives from intrinsic qualities of certain methodologies or the force of better argument. A convincing case for the production of 'better' (i.e. more valid or functional) knowledge will not *necessarily* give a methodology legitimacy and lead to its adoption.

The case study: integration of services for vulnerable adults in Lyndon

We turn now to the case study. As a single case linked to conceptual development, this is intended to enable the reader to 'see the world in a new way' (Siggelkow 2007: 23) through showing how Beetham's abstract concepts are manifested in the real world. Our claim here is for new middle-range theory (Merton 1967; Ling 2012), as we demonstrate the significance of legitimacy judgements in the practice of evaluating a complex intervention, and so in explaining the way an evaluation evolves as methodological choices are made. While the specifics are, of course, unique to the context, we are suggesting that similar processes may well be found elsewhere, particularly as the case exemplifies many contemporary challenges for policy evaluation: overlapping initiatives in complex, dynamic organisational settings, with insufficient resources committed for evaluation.

Between 2011 and 2013 the Lyndon Project aimed to integrate adult care services on a housing estate of some twelve thousand inhabitants, located on the periphery of an English city. The project's starting 'hypothesis' was that integrating place-based resources and services and focusing on building resilience, would improve service quality and so impact on the health, wellbeing and independence of older and vulnerable people (*Project Mandate, July 2011*)[4]. It would also reduce costs. Led by the local authority's adult social care commissioning team, it brought together council departments and organisations from the health and community sectors. The project comprised six workstreams[5], organised through a management group (headed by a 'project lead' assisted by the 'project manager') and overseen by a 'Project Board' of senior local authority and National Health Service (NHS) executives and mid-level managers of concurrent integration projects, chaired by a senior officer as 'project sponsor'.

From the outset the project had multiple functions. Alongside delivering substantive benefits to Lyndon's residents it was also a pilot, simultaneously learning how to integrate services and developing a business case for taking the approach forward elsewhere in the city. It was watched particularly closely by the managers of a much larger service integration project: *Getting It Right* was a partnership of the local authority and NHS which aimed to save £60 million annually through reducing unscheduled admissions of older people to hospital. Evaluation for both accountability and learning was thus at its core, but without dedicated resources apart from limited staff time. Neither was a methodology established at the outset: like other activities evaluation was experimental and emerged over time, and turned out to be complex and contested. This may well also be typical of myriad initiatives undertaken by local authorities and even at central government level, away from the well-resourced evaluation of flagship policies. It provided an environment in which policy makers' views on different evaluation methodologies became very visible, in the absence of an imposed evaluation framework and professional evaluators[6].

Our ability to observe arose from our position within the project. Formally this was based on an agreement between the local authority and our university, for us to 'support the [Lyndon Project] through … being a critical friend to the Project Board' (*Local Integrated Services Project Agreement, August 2011*). This gave both of us positions on the management group, and Connelly membership of the Project Board. Although we had no formal status as evaluators, in the context of the team's uncertainty over how to evaluate the project our

[4] To preserve anonymity, statements such as this one are not attributed to traceable sources. The titles are genuine to indicate the nature of the source.

[5] The 'workstreams' were Engagement & Communication, Market Development, Integrated Health & Social Care, Multi-agency Approach & Community Interventions, Intergenerational Intervention, and Evidence Base & Evaluation.

[6] The evaluation literature is unsurprisingly biased towards situations in which evaluators are involved, yet practitioners must be carrying out their own evaluations everywhere, all the time, even if they are not always labelled as such.

academic understanding of evaluation methodologies was increasingly drawn on. We had no authority to impose, but we advised, supported and informed as the practitioners developed their approach.

As researchers this allowed us to build up relationships of trust and friendship which gave us close access to the project (cf. Bjørkeng *et al.* 2009), spanning the two and a half years from the project's inception to the disbandment of the Board and the mainstreaming of some activities. For this paper we have drawn on digital recordings of: a set of eight interviews early in the project with the workstream leads, project lead, sponsor and manager; two sets of interviews with Board members (six from mid-project and thirteen towards the end); nine ad hoc meetings with the sponsor, project lead and Evaluation workstream lead; and almost all the steering group and management team meetings (thirty meetings altogether). We also have detailed notes from informal meetings and conversations with the lead and sponsor, along with field notes taken in and after meetings organised by associated projects within the authority and local NHS. The analysis also draws on the local authority's minutes of the project management group and Board meetings, and a number of project documents (referenced in the text e.g. the *Project Mandate* cited above.) Analysis of this material was done thematically, drawing broad themes from Beetham's framework and coding inductively within these (Braun and Clarke 2006). We were looking for enactment and discussion of rules, justifications of those rules, and for consent and dissent, including not only people's own practices and positions on these, but also their beliefs about and attitudes towards those of others.

Given the nature and purpose of this paper, a number of methodological points need to be acknowledged. Firstly, while this is not an evaluation of the project, there are clear parallels with a theory-driven evaluation in our analysis, in that we are presenting a theoretical explanation of a complex intervention. This reflects our own critical realist perspective, and our aim to develop 'practically adequate' causal accounts of complex processes, based in part on the interpretation of participants' accounts. Secondly, given our concern with how evaluation is presented, as well how it is carried out, we note our choice to present a narrative. Given the diversity of positions we uncovered, which points to the robustness of the methodology (Maxwell 2012), narrative is useful as it provides coherency based on causal links between events (Dodge et al. 2005; Abell 2009) and at the same time uses participants' understandings of how and why things were done to show how a project and its context are interwoven (Costantino and Greene 2003). We suggest that the validity of the analysis arises not just from the use of a diversity of sources, systematic coding and so on, but also in particular from our close engagement with the project. Our grasp of the situation was continuously checked and challenged by the practitioners, and its accuracy was evidenced by our engagement being sustained as the project was extended and by the way our views and advice were taken increasingly seriously.

Finally, the ethical dimensions of validity were closely considered. We were very open during the project about our dual role as critical friends and researchers, and our predisposition to theory-driven approaches. However, our ability to maintain a critical distance was challenged: we came to respect and like the practitioners, and appreciate the herculean nature of their task and the progress they made. Given that our task here is not to evaluate their work, but rather to examine the processes by which evaluation was shaped, this lack of distance is, we believe, relatively unimportant, and outweighed by the benefits of unusually good access to the field (cf. Patton 2015).

We present our account as a narrative in three chronological stages, through which we trace the evolution and relative 'fates' of (broadly) quantitative, qualitative and theory-driven approaches to evaluation. These stages follow the structure of the evaluation activities within the project:

A. an initial phase of working out how the project could and should be evaluated;
B. the execution of this, culminating in the evaluation report of the 'Winter Planning' activities; and,
C. the second-year extension of a slimmed-down project, ending with the final evaluation report.

At each stage we present a tabular summary of the legitimacy of contending methodologies in terms of Beetham's three dimensions of rules, justifying beliefs and consent.

The early months: how should the project be evaluated?

Early project team meetings exposed serious concerns about which evaluation methodologies would be acceptable within the team, and to other stakeholders, as ways of generating authoritative knowledge. This puzzling over legitimacy can usefully be summarised in terms of five questions:

- *Why* was the project to be evaluated? The team were unanimous that the key goal was to inform better service provision, but demonstrating cost neutrality or savings was also essential.
- *What* was to be evaluated? The whole project and its overall outcomes, the outcomes of specific component interventions, or the process?
- *Who* would evaluate, given that the allocated resource for evaluation was a single team member?
- *When* was it to be evaluated? Was there a need for a plan from the outset, or could the evaluation evolve with the project?
- Finally, *how* was it to be evaluated? Were numbers needed, and if so would these measure financial or other resource savings, or the health and wellbeing of residents? If not numbers, then what?

Underlying these was a deeper concern. In the context of funding cuts, the team feared that service integration might simply be used to reduce costs, rather than maintain or improve service quality. They were committed to achieving the latter, and so wanted an evaluation of quality as well as cost - yet achieving this appeared to them to be beset by a fundamental problem. They recognised that the accepted 'rules' for policy evaluation were to assess outcomes quantitatively, justifiable both as the usual, taken-for-granted correct process and by their specific need for a costed 'business case'. As the project sponsor put it, 'in crude terms what ‘The System’ needs to know is if the investment in the community-support workers is more than paid back’[7]. Yet the team also believed that showing relevant, significant changes in spending *and* attributing these conclusively to project activities was unlikely.

Conversely, from the outset they agreed that showing quick, positive results from specific interventions for individual service users would not be difficult, yet that despite the ‘immense power in the individual stories’ this would not be legitimate evidence in the eyes of the people who would be making decisions about resources. As the first Evidence Base & Evaluation workstream lead said to a project meeting,

> *we’ll have to do a lot of work with the Board, because I’m not sure they’ll accept [stories] as being enough, because how can you measure the cost-benefit of that, because it’s the value to the person?*

Beetham’s criterion is that rules should be justifiable ‘by reference to shared beliefs’. The problem for the project team was that although within the group they shared a belief in stories as evidence, they also believed that qualitative evaluation was not legitimate for other key stakeholders. Such beliefs were surprisingly widespread: some of those attributed (by others) with a need for numbers, including economists and statisticians, also expressed a belief in the value of qualitative evaluation. In turn they projected a need for ‘hard quantitative data’ onto service commissioners and, ultimately, the UK Treasury.

Although at this early stage we suggested that a causal approach might be useful, this was dismissed: it was unjustifiable according to the team’s existing beliefs about evaluation. In an early meeting the project manager bluntly declared that ‘if we’re honest this cause and effect thing is never going to be demonstrated’. Clearly our beliefs, and status as academics, carried little weight.

At the outset two basic positions thus emerged as candidate legitimate answers to the *what?* and *how?* questions: quantitative evaluation of overall outcomes, and qualitative accounts of individuals’ experiences. These were not conflicting ‘sides’ in a debate, nor was there any clear leadership to bring about a resolution – everyone seemed genuinely undecided, and committed to puzzling out the best evaluation methodology.

Table 1 summarises this early situation in terms of Beetham’s dimensions: both positions have claims to legitimacy *and* serious weaknesses, with neither able to establish dominance.

[7] Quotations have been used relatively sparsely in this text: in the nature of the discussions in the meetings and wide-ranging interviews there were rarely self-explanatory fragments of speech. We have selected phrases which are representative *and* intelligible to illustrate and reinforce the points being made.

Evaluation Approach	Rules	Justifying Principles	Consent
Quantitative evaluation of overall project by trained evaluator	Measures of costs, service provision, user satisfaction; applied at beginning and end of project	Principled: the validity of quantitative data as objective, scientific evidence Pragmatic: the need for evidence of cost savings	Explicit consent was high but grudging, tempered by scepticism about feasibility
Qualitative evaluation of separate workstreams by non-evaluator staff members	Narratives of immediate personal experiences, primarily of service users but also of project workers	Principled: reliability of accounts. Pragmatic: provide useful information on actual project impacts and of how these come about	Explicit consent was high, but heavily qualified by expectation of low legitimacy outside the group

Table 1. Early stages: competing methodologies

The *why?* question was addressed by making the evaluation explicitly multifunctional, using the device of a 'balanced scorecard' (Appendix 1), which positioned cost savings as outcomes alongside benefits to clients, operational change, and organisational learning. This approach was introduced by the project's most senior NHS officer, and carried the authority of use within the NHS, and its authorship by a recognised academic evaluation expert (see Moullin 2009). The scorecard could thus be adopted as defining a set of rules which both satisfied the team's belief that substantive outcomes mattered more than cost savings, and could also be expected to have external legitimacy. More or less simultaneously an organisational solution was devised which made it possible to avoid confronting directly the *who? what?* and *how?* problems. (*When?* was resolved by default, as the project started before evaluation was sorted out.) The solution was to split evaluation into two strands, carried out in different ways by different groups. This was set out in the project's *Overarching Evaluation Framework (April 2012):*

> *The Balanced Scorecard approach will be used to assess the overall impact of the numerous interventions in Lyndon. Each of the individual interventions will undertake their own evaluations independently.*

While the *Framework* defined the Scorecard and its indicators, it did not specify how the separate intervention evaluations were to be done, nor what indicators they should use.

The overall evaluation was assigned to the Evidence Base & Evaluation workstream, with a trained statistician as its new lead, and was thus separated from the five implementation workstreams. The task was technically challenging. In the absence of suitable neighbourhoods as controls, comparison of outcomes in Lyndon with the average for the city received widespread consent as a meaningful measure, in part because it was seen to partially solve the attribution problem. If it could be shown that since the project started

Lyndon outcomes were ‘better’ than the city overall, then it would seem reasonable to attribute this to the collective impact of the project interventions. This might tell decision makers that the Project *as a whole* was worth replicating elsewhere, though it was recognised that it ‘would be an analysis of the group effect rather than a cause and effect methodology for each of the individual activities’ (*Overview of Balanced Scorecard and Planned Work, April 2012*).

The middle stage: conviction without proof?

Designing the overall evaluation took many months. Meanwhile, the evaluation of the separate interventions developed in two parallel ways. While the puzzling continued over how to explicitly prove the worth of the project’s various components to external audiences, the team engaged in an extremely effective evaluative process. Never named as ‘evaluation’, it was a set of practices of reporting and learning which informed project development.

The explicit evaluation. The project’s managers went back to its hypothesis (Page XX above) and re-presented this as a matrix (Appendix 2) of 17 outcomes allocated to 15 activities, in order to generate a collective understanding of how the parts of the project related to the hypothesis. The matrix also provided a basis for monitoring progress, and so for evaluating it. In contrast, we (the authors) tried to use the hypothesis as the basis for explicit development of a ToC, identifying pathways from inputs to outcomes. In Beetham’s terms, although we had yet to work out rules for carrying out an evaluation based on this, the justifying principle was clear to us: the ‘standard argument’ for TDE that causal reasoning would show how process and outcomes were connected, and link individual interventions with the overall outcomes (Berriet-Solliec *et al.* 2014). Moreover, this approach was legitimised for us by its academic standing – something only we were aware of.

In a rare explicit discussion of competing methodologies, these two proposals were put to a team meeting. The matrix was introduced as a management and evaluation tool, legitimised by its transparency and simplicity, its grounding in the hypothesis, and by being proposed by the project managers. The Balanced Scorecard was defended by the Evaluation workstream lead on the grounds that it had been approved by the Project Board, and that overall outcomes could not be abandoned in favour of process measures. It was agreed that the two were complementary. Our theory-based approach was hardly discussed, and withdrawn on the grounds that adding another table was unhelpful. Its obvious (to us) superiority in evaluative power was insufficient to command the consent of the others: the rules were unfamiliar and comparatively complex, and academic standing insufficient on its own to legitimise our proposal. The enduring result was to embed a patchwork approach to evaluation, using a wide range of methods and a mix of qualitative and quantitative evidence, and eschewing causal attribution. Board meeting minutes, a few months later, note that

> *It is sometimes difficult when evaluating the whole programme to attribute which intervention/s met the outcome/s. The hypothesis is about doing a number of*

> *interventions at once to improve outcomes. All projects in the programme are being evaluated to ensure that there is evidence of their effectiveness. Very direct attribution is not always possible though.*

Yet, despite our private frustration, this approach was ultimately successful.

It was exemplified by the internal evaluation of the project's first intervention. 'Winter Planning' involved support workers visiting 'at risk' adults to discuss with them how they would cope in the event of a heavy snowfall. The evaluation report was based on the team's reflections, along with responses to structured interviews with clients. It was largely qualitative, with a single quantitative section on the impact on clients and their satisfaction. Throughout it was enlivened with quotations and vignettes of individuals' lives; a long, separate section presented individual case stories. These were not illustrative of a quantitative evaluative core: they *were* the core, with a rhetorical force exemplified by the report's front cover, which quoted 'Mrs K' as saying "Knowing somebody cares if you live or die during these dark, cold winter months means everything."

Tacit evaluation practices. These stories emerged from the most effective, but least visible, evaluation practices. The project management's concern with learning and improvement meant that every report by the support workers, and the discussions from every meeting, fed into a continuous, unstructured reflective examination and use of evidence. Some of this not only guided further action, but was also repeatedly reproduced to support the case for the project. This evidence was almost entirely stories about individuals in the community – of how they had responded to interventions by the project, or conversely of how the state had failed to intervene effectively in their lives (Box 1 paraphrases a typical, much-used example.)

These stories were re-told at every level from support worker supervisions through to city-scale design workshops for *Getting It Right*. They were emotionally powerful (Hoch 2006): funny, dismaying, reassuring, capable of moving senior professional audiences to public tears. The team were well aware of this power: as the project manager told the Board, "you can make the biggest impact with a story, a case, no matter how uninspired uninterested people are". While they were often shocking, the stories gained credibility through confirming the 'common sense' that services were currently inefficient and that change for the better was possible. The stories worked as practical evaluations: they convinced everyone involved that effective and cheap service integration could be achieved fairly straightforwardly, through

Box 1: A typical, much repeated story.

The local authority had removed a roadside bench in Lyndon. Unknowingly this effectively trapped a frail elderly couple in their house, as they couldn't take their daily walk to the daycare centre without a rest. The Lyndon Project community support workers discovered this; the multi-disciplinary team passed on the message within the local authority; the bench was replaced. Now the couple can again go and have their sausage roll at the centre – the only time they ever get out of their house.

community support workers linked to cross-sectoral teams, built up organically as needs were uncovered and trust was built between workers.

Strikingly, the stories' legitimacy as evidence for the success of this 'Lyndon way of working' went beyond the Project. Their repetition effectively shaped action within the much larger *Getting It Right* project, convincing both its chief executive and some general practitioners. The latter started taking on community support workers to start Lyndon-like programmes without any quantitative evidence: senior NHS managers were described to the Board as being 'quite nervous' that 'demand from GPs has overtaken the evaluation'. It also surprised the local authority staff, who stereotyped 'the bean counters in Health' with a highly problematic need for quantitative data. However, echoing Gawande (2010), *Getting it Right*'s chief executive suggested in an interview that the dominance of quantitative evaluation in the NHS was counterbalanced by a professional disposition amongst clinicians towards the evidential value of individual case histories. More bluntly the project lead credited the GPs with 'common bloody sense'.

At the end of the first year we can summarise the evaluation processes as shown in Table 2, with three methodologies sufficiently legitimate to be running in parallel, and TDE excluded.

Evaluation Approach	Rules	Justification	Consent
Assessing outcomes of individual interventions	Qualitative or quantitative assessment	Pragmatic: needed finer-grain evidence of impact than overall evaluation provided, and any kind of evidence seemed to be convincing Authority: preferred by project manager	High
The balanced scorecard	Quantitative evaluation using standard techniques	Principled: an accepted way to evaluate across multiple objectives Pragmatic: would generate quantitative outcomes data for the business case and service improvements Authority: sanctioned by NHS use and academic source	High in principle, but carried little/no weight in practice
Stories	Testimony of change from clients or front-line workers	Principled: truthfulness and authenticity of accounts Pragmatic: rhetorical effectiveness in convincing people the approach worked	High in practice, though constantly queried in discussion
Theory of change	Interview-based causal narratives	Principled and pragmatic: providing proof of 'what worked' required causal understanding Authority: academic status of TDE	Low, apart from the authors

Table 2: Parallel evaluation practices

The mature stage: a simple, effective project

During the middle stage little progress was made in several of the workstreams, and activity narrowed to developing the 'Lyndon way of working'. Senior managers decided to continue the project for a second year, focusing on this approach and its evaluation. The enduring legitimacy tension is clear. Continuation was based on the conviction of the Board and those they reported to that they had sufficient evidence that the approach was working. Yet they also believed that the evidence was insufficient to support a final decision about either long-term continuation or scaling up. For this they needed the crucial cost data to make 'the business case', particularly if *Getting It Right* were to adopt the approach formally.

The evaluation was developed within the project team and through informal meetings between ourselves and the managers, reflecting the trust that was developing and our increasing status as usefully critical friends. The resulting plan was for a multi-strand evaluation, comprising:

- quantitative measures of health and social care outcomes
- the collection of as many case stories as was practicable and analysing these for causal patterns
- an interview-based investigation of causal processes from the perspective of the steering group and project managers, conducted by ourselves.

Compared to the Winter Planning evaluation, there was significantly more emphasis on producing quantitative data, albeit rather grudgingly. As a senior officer put it, *'*I don't like that it's numbers but I understand, and it might be a way of shifting resources from health to social care.' TDE also gained some legitimacy. The intention to collect and analyse significant numbers of case stories, rather than just to use the most striking ones, was intended to develop robust and fine-grained causal accounts of how the project worked (Dodge *et al.* 2005). The project managers had clearly shifted their position on the value of such analyses, even if they would not commit to framing the evaluation around a comprehensive ToC.

At this point the quantitative, outcome indicators re-emerged in ways which challenged earlier negative assumptions about their feasibility and value. The Evaluation workstream produced figures showing that, since the project had started, Lyndon had bucked city trends for unscheduled hospital admissions. In response, concerned that this quantitative evidence was dominated by health outcomes (and despite her aversion to quantification) the project lead produced data showing the neighbourhood also bucking trends on the average age at which care packages were taken up. The latter was particularly valuable, as credible attribution (Mayne 2012) was easy, with a very short causal chain from *we did this for an individual* to *they didn't need a care package this year.* The financial implications were startling, as the high cost of each package implied huge savings, sufficient to pay for the community support workers many times over.

The project lead's final evaluation report contained most of the proposed elements, as well as the projected city-wide savings, despite her misgivings that these were 'too good to be true'. The systematic analysis of multiple stories was excluded. Time and resources ran out, and in any case she believed it was unnecessary to push the boundaries of locally-acceptable evaluation methods in this way. As the report put it,

> *the evaluation approach has been to collect a range of evidence to build up a picture of the change and impact of the work. No single piece of evidence is conclusive but the combined picture provides a body of evidence for commissioners to use to inform future decision making (Lyndon Way of Working, Evaluation, September 2013)*

The overall product was thus a legitimate patchwork, with different rules for different parts of the evaluation, justifiable by a corresponding range of principles. There was no single underlying rationality: or rather, the rationality was that of evaluation as craft. The project lead and her team had successfully constructed an 'intelligible picture' (Sanderson 2000: 449) (summarised in Table 3), using data of the right kinds, reported in an appropriate way, to do the required task of convincing an already sympathetic audience that the Lyndon way of working was a success.

Evaluation Approach	Rules	Justification	Consent
Multi-facetted, patchwork	Qualitative and quantitative assessment of outcomes; qualitative stories of process	Pragmatic: it worked! This was conclusive in the appropriate arenas, meeting both learning and accountability functions	High
Evidence-based, inductive 'Theory of Change'	Accumulating individual stories to become more than 'just anecdotes'; identifying patterns in these; linking to quantitative outcomes through interviews	Pragmatic and principled: every bit of data can be used as evidence and cumulatively they can tell a convincing story which explains the unexpected outcomes figures	Insufficient to be used, but rising amongst key staff

Table 3. The final report and the possible emergence of TDE

Conclusions

The overall story of evaluation in Lyndon is one of practitioners crafting a workable approach in the face of the indeterminate and competing legitimacy claims of different methodologies. At the outset there seemed to them no given, single set of rules for how to do an evaluation which would meet their needs for learning and accountability. Resolution lay in breaking the evaluation into methodologically and organisationally distinct parts, underpinned by different rationalities. These were then assembled into a multi-facetted approach which commanded sufficient external legitimacy for the ‘Lyndon way of working’ to be mainstreamed after the project’s end, and in part adopted across the city by the much larger *Getting It Right* programme. Reaching this point was supported by continuous, tacit, qualitative evaluation practices which enabled learning and development within the project while generating rhetorically powerful, convincing narratives for external audiences. Hearts and minds were being won over by informal practices at the same time as people struggled to formulate a publicly legitimate evaluation approach.

This struggle was one of engaged and reflexive practice, and significant shifts in what constituted accepted rules and justifying principles (Beetham 1991) took place over the project’s 30-month life. Central to the struggle was the tension between the legitimacy of designs based on familiar, accepted methodologies and legitimacy based on utility, with the former tending to reduce usability (Saunders 2012). Three aspects are of particular interest.

Firstly, quantitative evaluation had little traction until late in the project, when its utility was established. The Lyndon project team were not ‘everyday positivists’. They were sceptical about outcomes measures, and were well aware of the attribution problem created by complexity and of the problem of showing quantitative results in a city-scale comparison. Their rationality was more instinctively interpretivist, with ‘gut feelings’ for the power of narrative. However, they simultaneously subscribed to the view that everyday positivism *is* dominant, and so displaced the need for numbers onto Others. It thus remained crucial in creating legitimacy (Johnson *et al.* 2006).

Secondly, in practice the most powerful methodology was that which was believed to be weakest in public. Telling and retelling an individual story (provided it followed tacit rules of relevance and truthfulness) had *de facto* legitimacy as a way of evaluating the success of the project, with sufficient authority to be the basis for significant decisions - the antithesis of stereotypical evidence-based policy making.

Thirdly, practitioners’ response to the complexity of the real world was not to attempt to theorise it explicitly, and theory-based approaches had little legitimacy at first, echoing Sullivan and Stewart’s findings that ‘[i]n the field, discussion of “ToC” induced little interest’ (2006: 192). However, over time the Lyndon managers developed an explicit interest in causal theory, although this remained inessential to their craft: the patchwork approach was successful without recourse to theoretical underpinning.

How general might these findings be? The Lyndon Project is plausibly an exemplar of a fairly typical situation: of intelligent, experienced people working in a very complex and dynamic environment, trying both to deliver a complex intervention and evaluate it. Though one would expect different outcomes in different settings (for instance where practitioners were less averse to numbers) we suggest that the processes involved are likely to be similar elsewhere (Maxwell 2012: 141). There is clearly a need for further research to explore this further and corroborate our claims (Sayer 1992: 246), and in particular to clarify what is context-specific detail unique to this case, and what is more general.

In particular, using 'legitimacy' seems a fruitful new way (Siggelkow 2007) to understand in detail 'how' and 'why' evaluations evolve over time. Following Beetham (1991) in separating out the rules, their justifying principles and the consent they command, we can show how issues of use and usability are intimately and complexly connected (Saunders 2012). Legitimacy in evaluation is more than simply people believing in the utility of a particular methodology. It may rest on all of: the in-principle acceptability of an evaluation on methodological grounds (particularly if this is reinforced by authoritative external endorsement); whether it is done well and follows the relevant rules correctly; and on whether it is acceptable to the relevant audience(s) on the grounds that it provides useful, usable knowledge in that context. The approach also allows the different kinds of uses noted by Weiss (1979) to be analysed alongside each other. While an evaluation may be useful because it creates usable knowledge (Saunders 2012), other uses may be as or more important in justifying evaluation design in any given context – for instance the need to achieve wider legitimacy by producing numbers for external 'bean counters'.

This has practical implications for evaluators, who should pay attention to legitimacy, and be open minded on how it is conferred. Encouragingly, we suggest that the force of everyday positivism should not be overestimated, and that even where it is publically professed it may not be dominant in practice. Moreover, because legitimacy rests on all three dimensions of rules, justifiability and consent, it is not inherently stable. Change is probably (almost) always possible, and in particular we highlight the possibility for learning during the course of an evaluation. It may well be fruitful to promote such learning, through providing opportunities for experimentation and reflection.

Finally, we turn to the issue of whether an evaluation needs to be, or can be, underpinned by a single rationality, and what the role of CR-based TDE might be. The Lyndon case shows that such coherence is not a necessary condition for success, and suggests that a pragmatic patchwork of different methods is a feasible approach in this kind of context, being both effective and legitimate. Achieving a successful patchwork requires flexibility – the ability and willingness to respond to emerging knowledge, changing demands for knowledge and stakeholders' responses both to the substance and form of evaluation. Of course, this approach in itself must establish its legitimacy: while it worked in Lyndon it might well face legitimacy problems in other contexts. Thus it will need to be argued for, have its usefulness demonstrated, and itself be evaluated.

It was very evident in Lyndon, and presumably elsewhere, that evaluations are themselves interventions in complex and dynamic settings, and so subject to the same complexity: multiple goals, poorly defined boundaries, changing understandings and so on. Given this, the 'rules' of TDE – in particular establishing a ToC at the outset (Ling 2012) – reveal an incongruity between the approach's recognition of the complexity of policy making and its assumption that evaluators can organise an evaluation based in a single rationality.

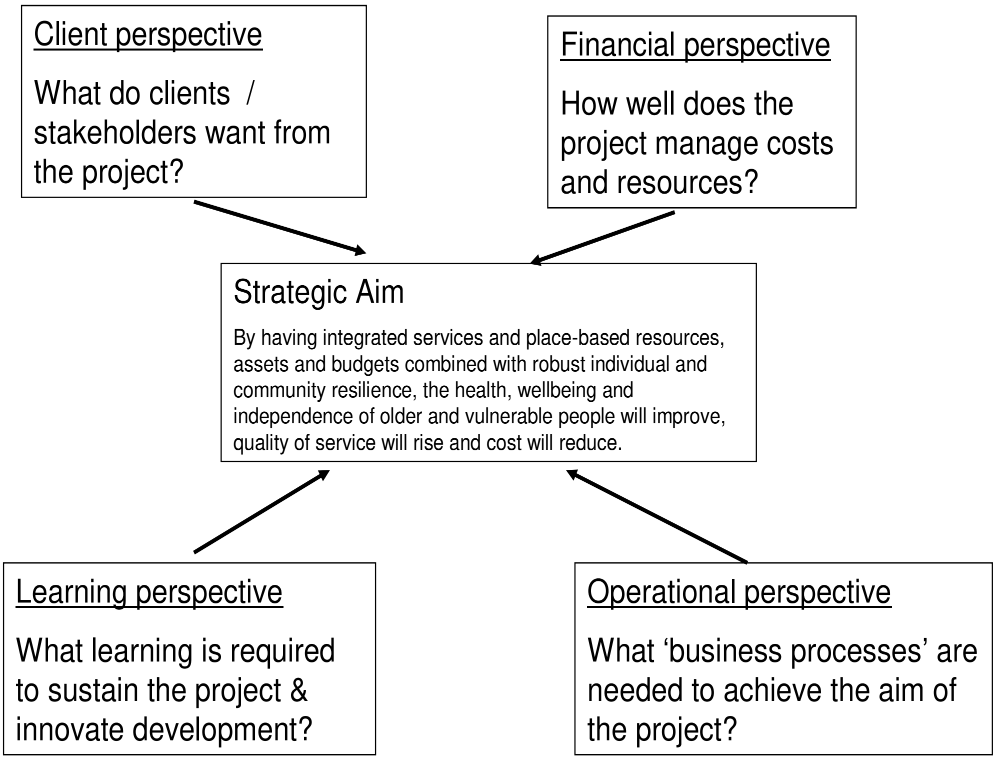
These points might appear to reinforce a turn to pragmatism as the overall rationality for evaluation as craft, as Sanderson (2000) seems to suggest. However, we end on a positive note from the perspective of those who, like us, believe that causal explanation is the bedrock of effective evaluation. On the one hand it is clear that causal theorising can be a legitimate component of a patchwork - where, for instance, short, unambiguous causal chains are present (cf.Sullivan and Stewart 2006, who note that practitioners often see the value of causal explanation.) On the other, more speculatively, we suggest that critical realist, theory-driven evaluation can act as a guide to the craft of developing successful patchworks – an underlying rationality which can incorporate different kinds of process and outcome data (Morton 2003; Connelly and Anderson 2007), and into which more explicit causal elements can be introduced if/when these gain legitimacy with an evaluation's audience(s). In such an approach, full Theories of Change act as regulatory ideals – not in themselves possible, but informing feasible, practically-adequate theorising. Our pragmatic goal should perhaps be to achieve evaluations which are more and increasingly theory-driven, rather than 100% theory-driven.

Funding

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

Appendix 1: *Lyndon Overarching Evaluation Framework*

The Balanced Scorecard approach will be used to assess the overall impact of the numerous interventions in Lyndon. Each of the individual interventions will undertake their own evaluations independently.



Client perspective

What do clients / stakeholders want from the project?

Financial perspective

How well does the project manage costs and resources?

Strategic Aim

By having integrated services and place-based resources, assets and budgets combined with robust individual and community resilience, the health, wellbeing and independence of older and vulnerable people will improve, quality of service will rise and cost will reduce.

Learning perspective

What learning is required to sustain the project & innovate development?

Operational perspective

What 'business processes' are needed to achieve the aim of the project?

Appendix 2 Lyndon planning matrix: the middle phase

	MULTI AGENCY, COMMUNITY INTERVENTIONS AND MARKET DEVELOPMENT				COMMUNICATIONS AND INFORMATION			INTERGENERATIONAL INTERVENTIONS		LBJ FORUM – COMMUNITY ENGAGEMENT					
Project work → **Issues/ other work** ↓	Resilience planning	Community money	Micro enterprise	Dementia strategy	Information capture and decision making	Local JSNA	Local directory	Intergenerational Interventions workshops	Review of pathways with view to joint screening	Support of smaller voluntary groups	Volunteering Including small scale community commissioning	Brokering/linking community groups	Channelling information	Sustainable business model	Prevention Pilots
Access to information	X		X				X						X	X	X
Getting around safely	X														X
Improved home care services		X		X		X									
Safety	X														
Reliable trades people, cleaners and volunteers			X				X			X	X			X	
Support with day to day	X	X	X								X				X
Lack of contact with neighbours	X		X					X							X
Key worker	X	X		X	X	X	X	X	X	X	X	X			X
Integrated services	X	X		X	X	X		X	X						
Resilience - community	X	X	X	X						X	X	X		X	X
Self care	X	X		X											X
Risk stratification	X				X	X									X
Whole household approach	X			X		X		X	X						
Theatre work								X							
Open spaces															
Neighbourhood appraisal/plan					X	X							X		
Learning and doing	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

FOR EACH PIECE OF WORK – NEED TO AGREE WHAT THE EVIDENCE BASE IS AND THE EVALUATION PROCESS. ALSO NEED TO ENSURE COMMUNITY INVOLVEMENT

References

Abell, P. (2009) 'A Case for Cases' *Sociological Methods & Research* **38**(1): 38-70.

Adelman, C. (1996) 'Anything Goes: Evaluation and Relativism' *Evaluation* **2**(3): 291-305.

Barnes, B. (2001) 'Practice as collective action' in T. R. Schatzki, K. Knorr Cetina and E. von Savigny, *The Practice Turn in Contemporary Theory* Abingdon, Oxon: Routledge, pp. 17-28.

Baxter, S., D. Chambers, M. Johnson, A. Sutton, E. Goyder and A. Booth (2017) 'Understanding new models of integrated care: a systematic review examining outcomes, pathways of change and applicability of evidence', at *Health Services Research Network Symposium*, Nottingham, 6-7 July 2017.

Beetham, D. (1991) *The Legitimation of Power*, Basingstoke: Macmillan.

Berriet-Solliec, M., P. Labarthe and C. Laurent (2014) 'Goals of evaluation and types of evidence' *Evaluation* **20**(2): 195-213.

Bjørkeng, K., S. Clegg and T. Pitsis (2009) 'Becoming (a) practice' *Management Learning* **40**(2): 145-159.

Braun, V. and V. Clarke (2006) 'Using thematic analysis in psychology' *Qualitative Research in Psychology* **3**(2): 77-101.

Chen, H. T. (1990) *Theory-driven Evaluations*, Thousand Oaks, CA: Sage.

Connelly, S. (2011) 'Constructing legitimacy in the new community governance' *Urban Studies* **48**(5): 929-946.

Connelly, S. and C. W. Anderson (2007) 'Studying water - reflections on the problems and possibilities for interdisciplinary working' *Interdisciplinary Studies Review* **32**(3): 213-220.

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

Costantino, T. E. and J. C. Greene (2003) 'Reflections on the Use of Narrative in Evaluation' *American Journal of Evaluation* **24**(1): 35-49.

Dodge, J., S. M. Ospina and E. G. Foldy (2005) 'Integrating Rigor and Relevance in Public Administration Scholarship: The Contribution of Narrative Inquiry' *Public Administration Review* **65**(3): 286-300.

Gawande, A. (2010) *Complications: A surgeon's notes on an imperfect science*, London: Profile Books.

Hoch, C. (2006) 'Emotions and planning' *Planning Theory & Practice* **7**(4): 367 - 382.

Højlund, S. (2014) 'Evaluation use in the organizational context – changing focus to improve theory' *Evaluation* **20**(1): 26-43.

Hurrelmann, A., S. Schneider and J. Steffek (2007) 'Introduction: Legitimacy in an Age of Global Politics' in A. Hurrelmann, S. Schneider and J. Steffek, *Legitimacy in an Age of Global Politics*, Basingstoke: Palgrave Macmillan, pp. 1-16.

Ivaldi, S., G. Scaratti and G. Nuti (2015) 'The practice of evaluation as an evaluation of practices' *Evaluation* **21**(4): 497-512.

Johnson, C., T. J. Dowd and C. L. Ridgeway (2006) 'Legitimacy as a social process' *Annual Review of Sociology* **32**: 53-78.

Lau, M. (2013) 'Flexibility with a Purpose: Constructing the Legitimacy of Spatial Governance Partnerships' *Urban Studies*.

Lehmann, E. R. (2015) 'What if ‘What Works’ doesn’t?' *Evaluation* **21**(2): 167-172.

Ling, T. (2012) 'Evaluating complex and unfolding interventions in real time' *Evaluation* **18**(1): 79-91.

Maxwell, J. A. (2012) *A Realist Approach for Qualitative Research*, Thousand Oaks, CA and London: SAGE Publications.

Mayne, J. (2012) 'Contribution analysis: Coming of age?' *Evaluation* **18**(3): 270-280.

Merton, R. K. (1967) *On Theoretical Sociology*, New York: Free Press.

Montague, S. and N. L. Porteous (2013) 'The case for including reach as a key element of program theory' *Evaluation and Program Planning* **36**(1): 177-183.

Morton, A. (2003) 'The theory of knowledge: saving epistemology from the epistemologists' in P. Clark and K. Hawley, *Philosophy of Science Today*, Oxford: Oxford University Press, pp. 39-58.

Moullin, M. (2009) 'Using the Public Sector Scorecard to measure and improve healthcare services ' *Nursing Management* **16**(5): 26-31.

Nutley, S. M., I. Walter and H. Davies (2007) *Using evidence: How research can inform public services*, Bristol: Policy Press.

O'Kane, R. H. T. (1993) 'Against Legitimacy' *Political Studies* **41**(3): 471-487.

Patton, M. Q. (2012) 'A utilization-focused approach to contribution analysis' *Evaluation* **18**(3): 364-377.

Patton, M. Q. (2015) 'Evaluation in the field: The need for site visit standards' *American Journal of Evaluation* **36**(4): 444-460.

Pawson, R. and N. Tilley (1997) *Realistic Evaluation*, London: SAGE.

Pouliot, V. (2014) 'Practice tracing' in A. Bennett and J. T. Checkel, *Process Tracing: From Metaphor to Analytic Tool*, Cambridge: Cambridge University Press, pp. 237-259.

Sanderson, I. (2000) 'Evaluation in Complex Policy Systems' *Evaluation* **6**(4): 433-454.

Sanderson, I. (2009) 'Intelligent policy making for a complex world: pragmatism, evidence and learning' *Political Studies* **57**(4): 699-719.

Saunders, M. (2012) 'The use and usability of evaluation outputs: A social practice approach' *Evaluation* **18**(4): 421-436.

Sayer, A. (1992) Method in Social Science: a Realist Approach (2nd edition), London: Routledge.

Schatzki, T. R. (2001) 'Practice theory ' in T. R. Schatzki, K. Knorr Cetina and E. von Savigny, *The Practice Turn in Contemporary Theory* Abingdon, Oxon: Routledge, pp. 1-14.

Schmitter, P. C. (2001) 'What is there to legitimize in the European Union … and how might this be accomplished?' in, Symposium: Mountain or Molehill? A critical appraisal of the Commission White Paper on Governance, Jean Monnet Working Paper No. 6/01, New York: NY University School of Law, pp.

Siggelkow, N. (2007) 'Persuasion with case studies' *Academy of Management Journal* **50**(1): 20-24.

Solmeyer, A. R. and N. Constance (2015) 'Unpacking the “black box” of social programs and policies: Introduction' *American Journal of Evaluation* **36**(4): 470-474.

Sullivan, H. (2011) '"Truth" junkies: using evaluation in UK public policy' *Policy & Politics* **39**(4): 499-512.

Sullivan, H. and M. Stewart (2006) 'Who owns the theory of change?' *Evaluation* **12**(2): 179-199.

Wagenaar, H. and S. D. N. Cook (2003) 'Understanding policy practices: action, dialectic and deliberation in policy analysis' in M. Hajer and H. Wagenaar, *Deliberative Policy Analysis: understanding governance in the network society*, Cambridge: Cambridge University Press, pp. 139-171.

Weber, M. (1968 [1922]) *Economy and Society*, New York: Bedminster Press.

Weiss, C. H. (1979) 'The many meanings of research utilization' *Public Administration Review* **39**(5): 426-431.

Weiss, R. S. and M. Rein (1970) 'The evaluation of broad-aim programs: Experimental design, its difficulties, and an alternative' *Administrative Science Quarterly*: 97-109.

Wistow, G. and H. Dickinson (2012) 'Integration: work still in progress' *Journal of Health Organization and Management* **26**(6): 676-684.

Zelditch, M. (2001) 'Theories of legitimacy' in J. T. Jost and B. Major, *The Psychology of Legitimacy: Emerging Perspectives on Ideology, Justice, and Intergroup Relations* Cambridge: Cambridge University Press, pp. 33-53.