Two concurrent but incompatible trends have emerged in development in recent years. Firstly, evidence-based policy and the results agenda have come become ubiquitous amongst government policymakers in recent years including in development. Secondly, there has been a realisation of the utility of systemic approaches to development policy and programming in order to bring about sustainable change for larger numbers of people. This paper highlights the negative impacts of this former trend on development and, more acutely, its incompatibility with the latter trend. The paper then highlights positive signs of a change in thinking in development that have begun to emerge to lead to a more pragmatic and contextually nuanced approach to measuring progress and identifies the need for further research in this area, calling for evaluation of approaches rather than searching for a silver bullet. The paper draws on a review of the evidence together with a number of key informant interviews with practitioners from the field.
“The Springfield Working Paper Series is aimed at improving the link between policy and practice in international development. These papers support the Springfield Centre’s wider role as an independent consulting, training and research organisation. Citation is permitted where properly referenced.”

1. Introduction

Evidence-based policy and the results agenda have assumed orthodoxy in recent years throughout the public sector and in the last decade this transition has been particularly rapid in the field of international development. A concurrent trend in development has been the recognition of the value in a more systemic approach to development issues; simultaneously addressing a complex range of interlinked problems to bring about sustainable change for greater numbers of poor people. The conflict that arises from the coexistence of these trends has yet to be identified nor addressed either in the literature or in practice.

Given this inherent conflict, this paper aims to formally identify these trends, the incompatibilities of their parallel employment, and set an agenda for addressing this conflict.

The paper contributes to an on-going discussion in international development policy and practice about the uses and misuses of results and evidence. It is divided into four main sections addressing two distinct gaps in the literature.

Section 2 begins by providing a brief history of evidence-based policy and the results agenda and how they are realised in international development, bringing together the identification of this trend in the literature across several fields.

Section 3 looks at how this has impacted on development policies, programme design and outcomes, which has yet to be identified in the literature, using the analogy of evidence-based medicine to illustrate the incompatibilities of the origins and current manifestations of the results agenda.

Section 4 follows by outlining the trend towards more cohesive, complementary development interventions aimed at sustainable and scalable change in a system; recognition of this trend in existing literature is limited. The particular ways in which the move towards a more prescriptive results agenda and evidence-based policy is inconsistent with this parallel trend are then examined, addressing a second clear absence of literature. Here, case studies are used to draw together the preceding sections of the paper in illustrating how programme design has been influenced by the evidence-based policy in systemic change programmes and the challenges such programmes have in compliance.

The concluding section then draws together some conclusions from the paper and suggests how these logical inconsistencies might be reconciled. The paper recognises the progress made in some areas by some agencies and calls for a research agenda covering how evaluation thinking and methodologies can be developed to capture systemic change.
The methodology used for the paper consists primarily of an examination of the literature but also partially through engagements in the field and with policy and programme development, including some case study data from key informant interviews. The combination of primary and secondary data helps in identifying trends but it should be noted that the degree to which these trends have been realised is by no means universal across development actors or subfields.

It is important first to define the meaning of and distinction between results and evidence. While the terminology is often fluid and confused, for the purposes of this paper, the key distinction is how the terms are employed. While results are used to assess the success or failure of given indicators for the assessment of performance by the actors involved, evidence is employed strategically to demonstrate the case for alternative forms of intervention. While there are historical differences in their employment and evolution, their analysis herein does not warrant a polarised appraisal as the implications of their introduction and prevalence for development has been similarly significant.

2. Evidence-Based Policy and the Results Agenda

Aspects of evidence-based policy can now be found across the public and private sectors, across academic disciplines and anywhere that management decisions have to be made and, indeed, evidence of evidence-based policy and the results agenda is far stronger than the evidence for it (Nutley et al., 2000; Black, 2001; Solesbury, 2001; Pawson, 2004b). New Labour’s 1997 manifesto slogan “what counts is what works”, which typifies the utilitarian turn in policy making, is far more evident in modern policy discourse and practice than examples of research counting whether ‘what works’ still ‘works’ and whether it works in other contexts. The slogan perpetuates perceptions of a silver bullet approach to public policy, belying the often complex systems underpinning social change.

2.1 Origins

Any perception that results and evidence as tools in management and planning are new phenomena is false. A long and cyclical history, formally identified in the literature as at least two centuries old, includes payment by results (PBR) being implemented in the mid-19th Century in state schools in England in order to improve learning outcomes (Rapple, 1994).

---

1 Indicators are variables in which the programme intends to affect change.
2 Interventions are the sets of activities conducted by development programmes.
The dominant discourse of the time included the recently reinvigorated term Value for Money (VfM) but the use of the tool in the UK had been widely discredited by the end of the century due to the increased bureaucratic burden and the failure to improve efficiency (Madaus et al., 1987; Eyben, 2012). Evidence too, sees its origins at this time in the field of experimental medicine (Bernard, 1957), but ‘evidence-based medicine’, and its tool of choice - the Randomised Controlled Trial (RCT), were not developed until the 1970s (Sackett et al., 1996).

The 1960s saw a revival of the results agenda in Anglophone countries reflective of a broader positivist turn in the social sciences in the guise of Management by Objectives (MBO), adopted by the US government along with the associated tools and terminology of Cost Benefit Analysis (CBA) and risk assessment (Espeland, 1997; Johnsen, 2005).

The emergence of New Public Management (NPM) in the 1970s throughout the developed world remains a paradigm of public sector management whose influence, discourse and practices have shifted, strengthened and weakened throughout that time (Hood, 1991; Hopwood and Miller, 1994; Gray and Jenkins, 1995; Power, 1997; Hood and Peters, 2004). The global political polarisation of the 1970s and 1980s meant that the influence of ‘evidence’ in policy was secondary to the influence of ideology. It did not matter whether trade with certain countries would have contributed to greater growth if it would have occurred at the expense of what was right. The Anglo-American shift to the left of the 1990s, together with the end of the cold-war resulted in the promotion of a perceived objectivity and pragmatism in policy-making, of which evidence-based policy was a major part (Black, 2001; Solesbury, 2001; Parsons, 2002; Sanderson, 2003; Pawson, 2004a). Since this reinvigoration of NPM in the mid-1990s, the trend has both increased in its strength of application within fields and diversified in its application to new fields of public policy (Ferlie, 1996; Pawson, 2004b; Haque, 2007; Lapsley, 2009).

2.2 Rationale

The rationales for results and evidence are clear and distinct. Results are used to ensure that people, programmes, assets etc. are performing to the required level in a defined and measurable way. Evidence is used to ensure that empirically derived knowledge is obtained on which of a range of alternative strategies for performing a given task is most effective, according to a pre-defined set of objectives, so that this strategy can be replicated and if necessary adapted. The definitions of purpose for both results and evidence are sound. Contention arises when one considers what qualifies as a valid form of evidence, an indicator of results and what, precisely, evidence should try to assess.
In summary, there are three primary reasons for the collection of evidence and results. Firstly, **accountability** – while efficient markets hold the private sector to account for performance, a lack of information on performance may result in inefficiencies in the public sector (Minogue et al., 1998; Barzelay, 2001; Lapsley, 2009). Secondly, **efficacy** – given limited resources, it is important to know where money can be spent to ‘do the most good’ (Bovaird and Davies, 2011). Thirdly, **learning** – determining the impacts of interventions so that lessons can be transferred to alternative contexts (Feinstein and Picciotto, 2000; Hall et al., 2003).

In medicine, from where the drive towards evidence-based policy originated in the 1970s, the population to whom the evidence is to be applied is biologically similar if not homogenous. Even when variation occurs, it is possible to state that during trials, the treatment was x percent less likely to result in a positive outcome for patients with a given genetic characteristic. The ‘treatment’ too is universal; each case will receive the same dose of the same medicine, delivered in the same way. Lessons learned from experiences with one group of people in one set of RCTs are likely to be largely applicable to another, provided the sample is representative.

The political shift of the 1990s saw the application of such methods spread toward social policy with the theory that behavioural change, taken across a large enough sample, was equally homogenous as biological change.

A key differentiation between results and evidence in rationale is in the principles underlying how they are employed which relates strongly to Goodhart’s law i.e. once a measure becomes a target it ceases to be a good measure (Elton, 2004). The rationale behind evidence relies heavily on performance being accurately measured and not used as a target whereas the rationale behind the results agenda relies on targets being set and then performance being measured against them.

### 2.3 Current Features in International Development

In international development, the transition towards evidence-based policy came later than in other public sector fields. Aid budgets increased dramatically through the 1990s and into the new millennium (OECD, 2013). With increased funding came increased scrutiny and, in the wider context of NPM, this meant international development became a more transparent, measured and target-driven sector. A change in public perceptions of aid was perhaps the strongest driver behind the change in donors’ approach to aid. International aid in the 1980s was characterised as altruism and compassion which fulfilled domestic political objectives of
donor governments. Some influential critiques (Easterly, 2002; Easterly, 2006; Clark, 2008) led some to question the efficacy of aid in reducing poverty while the global financial crisis led to questions of its affordability (Te Velde et al., 2008; Dang et al., 2009).

High level OECD meetings on aid effectiveness were held in Rome (2002) and Paris (2005) which led to The Paris Declaration on Aid Effectiveness of 2005. These were followed by meetings in Accra (2008) and Busan (2011). The declaration is based on five key principles:

i) Ownership of development priorities by recipient countries
ii) Alignment of donor funding behind these priorities
iii) Coordination between donors to avoid duplication
iv) Focus on the results of aid which should be measured and used as a management tool
v) Transparency and accountability between donor and recipient countries (Rogerson, 2005; Droop et al., 2008).

The first two of these points should have meant a shift towards direct budgetary support while the final two of these principles are of particular importance to the way evidence-based policy has increased in prominence since the Paris Declaration (De Renzio, 2006). The fourth principle formalised the introduction of results-based management (RBM), which had obtained orthodoxy across the public sectors of Northern countries throughout the 1990s, to the field of development (Strathern, 2000).

The practical ways in which this shift to an accountability culture has manifested itself in development are evident from budget allocations and programme design right through to the micro-management of interventions and reporting. At a programme design level the trend has worked in parallel to the role back of the state to mean that the number of directly contracted staff in major donor organisations has been greatly reduced (Easterly and Pfutze, 2008; ICAI, 2012). The result of these mutually reinforcing trends has been a proliferation of outsourcing of all stages of the development process with programme design, implementation and evaluation now all put out to competitive tender. NGOs, development consultancies and major international engineering and accountancy firms now compete for contracts as has occurred in other areas of the public sectors including the health service and defence (Domberger and Rimmer, 1994; Stubbs, 2003; Altenburg, 2005; Harland et al., 2005). When combined with the, albeit wavering, general budgetary support to developing

---

3 Despite featuring heavily in the Paris Declaration, direct budgetary support has not increased as one might have expected in recent years due to concerns over good governance together with the use of
country governments that emerged from the Paris Declaration, the direct control of donor staff over the trajectory and success of programmes has been greatly reduced at a time when the pressure for better results is increasing still further. Donor staff levels have been greatly reduced as has the technical capacity within them (DAC, 2010), while the budgets they have control over have increased rapidly (Mehrotra, 2009). Indeed, the global financial crisis had the counter intuitive effect of increasing aid from $104bn in 2007 to $134bn in 2011 (OECD, 2013). It is perhaps inevitable, then, that development has succumbed to the trend of management which is rigidly dependent on quantitative targets.

While the Paris Declaration was partially responsible for both the results agenda and the drive towards evidence-based policy, the ways in which they have manifested themselves in development have been quite different. The results agenda impacts on programme level dynamics whereas ‘evidence’ is intended to inform policy across programmes. Current features of the results agenda include the proliferation of Key Performance Indicators (KPIs) - which are employed from the level of the individual employees through to contract management and payment by results – also variously called ‘payment for success’, ‘milestone payments’, and ‘performance-based objectives’ – which involves a proportion of the contract fee being put at risk against the production of these numbers. Indeed in a recent business case for an innovative programme in West Africa, the donor confessed that while they were fully aware that these simplistic output targets would not help them achieve their broader goals, they were obliged to make payment contingent upon them (Key informant interview with programme manager). These are the management tools identified in the fourth point of the Paris Declaration. As part of the DfID Business Plan 2010-2015, the department now publish figures on 15 quantitative measures of development performance, all of which involve a number of people, an amount of money spent or the number of items purchased.

programmes and, here, it is the RCT, together with other experimental and quasi-
experimental methodologies, that form the primary weapons in the armoury of those that
seek to compare efficacy and impact across development programmes. Indeed, an implicit
hierarchy of evidence has emerged. For people – those few within donor organisations
charged with making such comparisons - with limited time and expertise in a given
geographical or thematic context, numbers, once again, have obtained primacy. Qualitative
data is not considered to be evidence of change but, rather, an explanation the change that
the numbers provide evidence of (Kusek and Rist, 2004). However, this is by no means
unique to development. For example, in the trialling of a new drug, patients’ reporting of
adverse reactions to a newly trialled drug was not considered significant as it could not be
detected quantitatively (van Grootheest and de Jong-van den Berg, 2004). Given an
ethnographic account of how systems altered so that farmers can react better to changes in
input availability or a study which compares two programmes and states, with statistical
accuracy, that a particular intervention improved incomes by more than another, it is easy to
see why such a hierarchy has emerged in a resource constrained environment.

Another practical manifestation which transcends both evidence and results is the ‘logical
framework’ or logframe, which, while it has been part of development practice for several
decades, has mutated in how it is used in development practice. Originally adapted from the
US military, the intention behind the logframe is to allow a theoretical causal chain to be
drawn between the activities that a programme undertakes, the expected outcomes of those
activities (monitoring and results) and the impact that those outcomes have on poverty. In
practice the use of the logframe has constrained adaptability and innovation in development
programmes as is explored below (Chambers, 1997; Gasper, 1997; Hummelbrunner, 2010).

An additional example of the way the results agenda is currently realised actually emerged
from practitioners as part of a drive to improve practice. Practitioners in enterprise
development have attempted to accommodate the results agenda is through the creation of
and adherence to the Donor Committee for Enterprise Development (DCED) Standard for
Results Measurement. This is a tool for both donors and practitioners to provide formalised
advice on best practice which is acceptable to both donors and practitioners (Thomas,
2008).

The UK’s Department for International Development (DfID)’s recently produced policy on
evaluation states:

Some development practitioners and researchers have promoted impact evaluation
through experimental methods and randomised control trials as carried out in
medicine. We recognise the usefulness of this type of work and support an increase in
rigorous impact evaluations more generally, but as one tool in the evaluation tool box, which must sit alongside evidence gathered through other evaluation methodologies (DfID, 2009; 14)

However, in the terms of reference for a recently commissioned impact evaluation of a market development programme it was stated that the evaluators will ‘use experimental and quasi-experimental methods especially randomised control trials wherever possible’ (Interview with evaluation programme manager).

Indeed, a 2012 DfID funded working paper (Stern et al., 2012) examined this very issue and concluded that “Appropriate IE [impact evaluation] designs should match the evaluation question being asked and the attributes of the programme” (79), calling for a more nuanced and context specific approach to the classification, use of and demands for evidence. Still, similar stipulations as to the necessity of experimental and quasi-experimental methods persist in the terms of reference for a wide range of bilateral and multi-lateral donors.

The increased financial scrutiny that accompanies a broader public sector resource constrained environment impacts further on the desire for comparability across development programmes as characterised by the re-emergence of VfM objectives.

Regular and effective monitoring, reviewing and lesson learning are key to how DFID measures the Results of its projects and demonstrates Value for Money (DfID, 2011; 2)

All DfID programmes are urged to ‘demonstrate value for money’ (Eyben, 2012). Where objectives and budgets are defined and payments are measured against them it seems odd to employ the phrase VfM in this context. If the value is the deliverable, be it an output such as a report or an impact such as the number of people whose incomes have increased, and the money is the budget as defined at the outset of the programme then it seems that VfM is a by-product rather than an objective of the ‘management by numbers’ approach.

A further point of note is that this drive towards accountability has not stopped at the donor level. The recently created Independent Commission for Aid Impact (ICAI) is the “independent body responsible for scrutiny of UK aid”. The centrality of evidence-based policy and the results agenda is evident throughout this body’s work as they “[C]hampion the use of independent evidence, to help the UK spend aid on what works best”. While the ICAI too is held accountable, in its case to the International Development Select Committee who in turn are accountable to their constituents, there is an interesting paradox in that the ICAI

4 Here, and in the remainder of the document, empirical observations from development programmes have been anonymised so as to protect the interests of key informants.
states on its website that it ‘accounts for its own performance’ while ‘championing independent evidence’ (ICAI, 2013). Numerical targets are also in evidence with the ICAI which seeks to publish 10 to 15 reports a year on its website.

### 3. Implications for Development Programming and Outcomes

The emergence of evidence-based policy and the results agenda together with contemporary management tools have impacted on development programme design and implementation and, ultimately, outcomes for poor people significantly. The changes in the international development sector were implemented with the explicit objectives of transparency, accountability and improved development outcomes relative to investment. The consequences of how these laudable objectives were implemented, however, have been far less positive. If development is viewed as a linear chain of power relationships then recent transitions have had the impact of shifting power gradually up the chain away from those ‘doing development work’ in developing countries.

The global organisation of aid and how it is organised and distributed varies by country and across time. Vielajus et al. (2009) provide a good summary of how development agencies are organised, and who they are accountable to, in the UK, Germany, France and Sweden. However, since this time, for example, the ICAI has been set up as an additional actor in the accountability chain beyond ordinary democratic parliamentary procedures. Similarly, USAID added a new division in 2011 entitled the Partner Compliance and Oversight Division whose objective was to monitor the performance of contractors.

Figure 1 represents a stylised version of the accountability trail of the international development sector in the UK as it now stands. As the results agenda has proceeded, the weighting of risks undertaken by each party has been pushed downwards. DfID are forced to report on the success of their work according to a narrowly defined set of indicators. They increasingly make contractors’ payments contingent on the achievement of the number which they consider acceptable and the contractors then make a proportion of their (often freelance) employees’ salaries contingent upon satisfaction of these criteria. Of course, the primary way to
minimise this risk for all parties is if programmes set modest, achievable targets that are easily measureable, static and clearly defined.

At a programme design level, then, the combined effect of the above drive towards evidence-based policy and the results agenda is that programmes are designed where the implementer of the programme has maximum control over the programme outputs.

This is by no means an argument that contractors in development are a wronged party in this new configuration of accountability. There are two notable sets of actors that are entirely absent from Figure 1. Firstly, developing country governments, who now, despite the Paris Declaration, are sometimes seen as an uncontrollable variable and therefore are bypassed in development programming wherever there is no explicit requirement for their inclusion, such as in state building programmes. If a seven-year development programme is set the target of increasing access to road networks for rural communities, the contractor is unlikely to take the risk of working through the government to identify impediments to access and build capacities if their payment is contingent on it and will, instead, simply build the roads themselves. Most importantly however, the people omitted from the current structure of development accountability are the intended beneficiaries; poor men and women. Evidence-based policy and the results agenda does not frequently account for the genuine impacts development programmes have on the lives of the poor. This refers not just to the minority that are ‘reached’ by development programmes but also to the vast majority of poor people whose lives are not influenced by this direct transferral of goods and services.

3.1 Comparing Apples and Existentialism

Central to the growth of evidence-based policy has been an assumption that methods drawn from natural sciences can be applied with equal efficacy in development. However, the conditions of development are entirely incompatible with those in natural sciences – taking natural science methods for use in development is to misappropriate them; to believe that ‘apples can be compared with existentialism’. A direct comparison of medical science and development illustrates this point.

Evidence-based policy, the results agenda and development programming have become increasingly incompatible. In medicine, a professor will run a laboratory with multiple researchers operating under their general mandate. Research on ‘what works’ begins at the

---

5 Many programmes – for example in health and education - are still delivered directly through developing country governments but none of the actors in the chain are held accountable to them, nor the recipients of the aid, for the programme’s results.
theoretical level and is then expanded to the cell level, conducted in live trials on mice before eventually very limited human trials can begin. Results are analysed including considerations of cost effectiveness, before eventually, a drug will be classified as safe. Even still, drugs can fail phase II or phase III trials if side effects are witnessed which impact on the continued use and uptake of the treatment. If one leaves aside rather fundamental issues identified above of the differences between development and medicine of plurality/homogeneity of ‘problems’ (illnesses/poverty causes), ‘treatments’ (drugs/incremental multifaceted social, political and economic interventions), and ‘treatment groups’ (humans/poor people, societies, communities, industries etc.) there remains a critical difference in how these two fields are approached in the current environment which impacts on the relevance of evidence-based policy. At any one time there might be hundreds of researchers around the world working on a daily basis on how a given drug impacts on the behaviour of a G-protein coupled receptor and its downstream signalling effectors, which in turn affect the likelihood of malignant transformation of cells. A good result may not be a cure for cancer but a positive indication of the areas in which treatments may be developed with a great deal of further research to address some forms of cancer in some people. It is on this data which evidence-based medicine is founded. Conversely development programmes are evaluated for their ability to, and compared in relation to, far higher level objectives, with far fewer examples on which to draw.

As a development comparison, one of the more highly specified systematic reviews examines the economic impact of conditional cash transfer (CCT) programmes (Kabeer et al., 2012). It bases conclusions on 11 programmes in nine countries. All bar one of the studies are of programmes in Latin America. The common problems exist of the wide range of differences in the programmes being compared in terms of their target variables, target treatment groups, years of operation etc. The majority of the conclusions are intuitive, such as the fact that CCTs are more effective in reducing child labour for older, male children as it is they who are more likely to have been in work prior to the CCT. Where conclusions were not intuitive, they were found to be highly geographically differentiated:

...we find that the impact of transfers on adult labour supply varied by gender, size and duration of transfer and type of employment (Kabeer et al., 2012; 20-21)

This conclusion is made despite 10 of the 11 programmes being based in Latin America, giving a certain degree of cultural comparability. There are also areas where the conclusions of the review are far more dubious when used to influence policy. For example, in the area of household consumption, a difference is suggested between in how female and male recipients use the CCT, with transfers to females resulting in higher quality nutritional intake
and increased purchases of children's clothing. Ultimately, despite all of the caveats contained in the document the conclusion is drawn that:

*CCTs appear to be an effective measure for achieving what they were designed to achieve: promoting children’s education and reducing child labour among poor and marginalised groups* (Kabeer et al., 2012; 44)

It is in the utilisation of such conclusions that policy is influenced. The source for this particular strongly worded conclusion is a set of 15 studies across seven countries, six of which are in Latin America⁶, with a range of impact variables and methodologies. This is the medical equivalent of stating that ‘chemotherapy cures cancer’ based on a research from a single lab based only on animal testing.

In the medical scenario it is feasible that one million petri dishes, ten thousand mice, and a systematic review of ten randomised controlled trials each featuring 100 patients together with thorough cost benefit analysis might go into the approval of a given treatment for one type of cancer with caveats that, for example, it should not be used given certain comorbidities etc. In development, the ‘solutions’ being sought are the equivalent of a panacea for all cancers or, arguably, something with a greater number of context specific causal factors. The number of ‘observations’ being used to make these generalisations represent an incalculably small fraction of those used in medicine. The causes or extent of poverty are not binary, and nor are the solutions. If a smoker develops lung cancer, stopping smoking won’t cure the disease but if a poor person is given the opportunity for well-paid employment then the likelihood is that they will be able to take themselves out of poverty.

Social problems are complex and varied and while some in development are recognising this complexity, reconciling this with the drive towards accountability and evidence-based policy has yet to be accomplished, as is explored in the following section.

4. Systemic Change, Evidence-based Policy and the Results Agenda

Official Development Assistance (ODA) totalled $133.5bn in 2011 with approximately 2.5bn people living on less that $2 per day (OECD, 2012). So, at current levels, if aid is regarded as directly delivering benefits/resources, aid could offer under $0.15 to each poor person

---

⁶ This is a subset of the broader systematic review paper referring only to the effect of CCTs on child labour and education rather than the systematic reviews which addresses CCTs as a whole.
each day. In the context of rising global food prices as population pressure increases, this would barely make an impact on the majority of people’s lives. Therefore, for aid to be effective, it has to stimulate wider change and there is a need to leverage current levels of aid to produce greater outcomes. There has been an increasing recognition in recent years that this is not likely to happen by through traditional direct delivery of ‘charity’ (Pronk, 2001; Rogerson, 2011). Instead donors have begun, discursively at least, to incorporate objectives of systemic change into a great number of their programmes in an increasing range of sectors.

Systemic Change, *Making Market Work for the Poor (M4P)*, global production networks and complexity science are all ways of understanding development processes as complex adaptive systems rather than linear input-output models. Since the turn of the millennium, there has been an increasing recognition of the complexity of development problems amongst donors and an associated move towards the incorporation of systemic change objectives into development programmes (see for example USAID’s Feed the Future, AUSAID’s PRISMA, and the multi-donor Katalyst programme). Systemic change was conceived of as a way to leverage aid inputs through emulation utilising local surpluses in the recipient country. Thorough analysis of the incentives of different actors within a system to alter their behaviour resulting in better outcomes for poor people allows development programmes to intervene only where this change is plausible. Then, in the absence of intervention, incentives remain aligned with pro-poor outcomes making changes in a system sustainable. Scale is achieved through demonstration, imitation and adaptation (Elliot et al., 2008). The potential for systemic change to deliver lasting pro-poor outcomes to large number of people has been recognised in an increasing range of fields including health, education and finance (Bloom et al., 2012; Peters et al., 2012; Ledgerwood et al., 2013)

While the discourse surrounding these trends has been changed over time, their popularity has been on an increasing, if undulating, trajectory. As discourse became more unified behind systemic change, particularly in the markets for services, following the influential World Development Report (World Bank, 2004) and a number of other relevant publications (Bear et al., 2003; Elliott and Gibson, 2004; Ferrand et al., 2004) the evidence agenda continued to gather pace as identified above. As early as 2002 the International Labour Organisation (ILO) (McVay, 2002) was simultaneously promoting the ‘systemic changes’ brought about by linking marginalised business owners with the mainstream economy while also employing the exclusively quantitative and reductionist *Performance Measurement for Business Development Services to Micro and Small Enterprises (PMF)* (McVay, 1999). More large programmes with systemic change objectives were launched and, with each iteration,
increasingly tightly defined measurement practices were identified. The subscription to the logic of systemic change was clear.

4.1 Incompatibilities between Evidence-Based Policy and Systemic Change

The drive for evidence-based policy and the results agenda and the drive for systemic change have occurred in parallel and the incompatibilities of these agendas have yet to be reconciled. This is perhaps reflective of a broader trend that donors have become more demanding as budgets and consequently scrutiny has increased in international development. As political concerns gain public support, they are quickly added to the specifications of what donors want. The climate change agenda means every intervention must be climate neutral/positive (Klein et al., 2007). The gender agenda means every programme must disproportionately benefit women (Kabeer, 2003; Moser, 2005). The fragmented nature of research, communications and policy divisions within donors results in a corporate cognitive dissonance meaning that this wish list grows ever longer with no conception of whether the concurrent satisfaction of these goals is possible or desirable.

There is a clear disconnect between NPM, which seeks to set up front numerical targets for the duration of a programme and manage through their enforcement, and a systemic change approach which seeks to implement short feedback loops and bring about institutional and behavioural change through detailed and ongoing analysis of local contexts leading to iterative and bespoke solutions which adapt over time (Blackman, 2001).

A systems understanding of development represents a typical example of the complex unbounded problem (Dovers and Handmer, 2012) which makes assessment through the rigid framework of NPM and its tools highly challenging. One of the primary issues is that there is no shared agreement about the problem (Harrison, 2000).

These are problems associated with setting up-front, quantitative targets in situations which need to be able to adapt and respond as the context becomes clearer and external circumstances change. Ultimately, such aspects of the results agenda address two of the three elements of its rationale. Contractors are held accountable for how well they execute the tasks which they have been set, and the results agenda performs this task well. However, it cannot address whether the contractors are being held accountable for the

7 The psychological phenomenon of believing in two contradictory ideas equally and simultaneously.
things most likely to achieve their ultimate goals. In monitoring, and indeed paying for, programmes in this way, implementers have no scope to deviate from what was agreed at the start of the programme, whatever the developmental implications of this might be. Secondly, the results agenda addresses issues of efficacy. It measures which of the predefined strategies resulted in the greatest change in the predefined indicators. What it does not allow for, in the monitoring of these programmes, is scope for the measurement of any additional indicators using alternative methodologies; capturing the externalities.

Perhaps the most significant problem with evidence-based policy comes in using the evidence gathered from these activities for the learning aspect of the rationale. Systemic change programmes are, above all, context specific. If a programme that had worked with a fertiliser producer in Nigeria to deliver training in usage at the point of sale was shown to have increased yields more than any other programme with the same goals, verified by an RCT, does not mean to say that this would be the most successful strategy in another part of Nigeria let alone in other developing countries. Social, economic, cultural and political circumstances dictate outcomes and it is the iterative and flexible nature of systemic change programmes that allows them to succeed. Moreover, in an intervention such as this, working through partners is highly dependent on personalities and the nature of the companies or organisations that interventions are implemented through. The sustainability and scale objective of systemic change programmes is reliant upon strong and capable partners that, by the end of the programme period, will be able to maintain and augment the change brought about by the development programme’s intervention in the absence of external assistance. The absence of such partners in an alternative context means that different ways of addressing the same problem may have to be employed.

In addition to alternative solutions, the very nature of the problem may also differ in alternative contexts. Low use of fertiliser in a different area may be due to low capital availability and a lack of micro-credit facilities as one of a wide range of other possible causal factors. Currently, the entire discourse characterises societies and the humans of which they are comprised as a homogenous mass with identical behaviours and responses. Where the incentives that underlie failures in systems are considered, they too are considered to be universal despite countless examples to the contrary.

What follows is a set of four examples to serve as case studies of the incompatibility of systemic change with how evidence-based policy and the results agenda are currently implemented; two based at the results level and two at the evidence level:
4.1.1 Independent Contractors and Online Databases

In one programme with systemic change objectives, a constraint on the system in which independent building contractors were operating was that small scale producers were not linked with the buyers where there was nascent demand. As part of the need to identify targets during the design process, one intervention was identified whereby a database would be created to overcome this constraint. The theory was that buyers simply didn’t know about the quality and volume of products and services available and so the creation of this database would lead to better linkages and increased incomes. A target was set to register 165 such contractors on the database; a target which was subsequently exceeded substantially by the implementer, scoring them highly in the annual review process.

Unfortunately, the review found that there had been no impact on poverty as no buyers used the database, internet connection and computers were absent and the construction workers did not see its purpose. This is, therefore, an example of where the results agenda has produced a false positive. The design team had designed an intervention in which there was no defined exit strategy and little prospect for the intervention to continue beyond the period of donor support. However, having previously identified a criterion on which the programme would be assessed which was not reflective of its broader strategic goals, the programme was bound to succeed by its own measures (DfID, 2012).

4.1.2 Thin Markets, ‘Thick’ Programmes?

In a small country in Asia, one systemic change programme was set the target by the donor of beginning a fixed number of interventions by the end of the first year. In the beginning of the implementation phase it was clear that initial market analysis in the design phase had been inadequate, meaning many of the areas identified for intervention lacked the appropriate partners to work with and some sectors would not yield rapid results in a systemic change programme at all and so should not be priorities. Due to the initial targets set, however, the programme was required to ‘start’ interventions. At the expense of working with partners in sectors with strong potential, staff were required to begin to produce compliance documents and begin working in industries in which they saw no potential for impact. The meaning of ‘starting’ in this context was not defined and so tokenistic efforts were made at fulfilling this goal although there was no indication of progress toward poverty reduction. This programme also scored highly in its annual review (Interview with Programme Director).
4.1.3 Where's the Evidence?

At the level of evidence on which to base policy, one development consulting firm was set the task of conducting a significant multi-year evaluation of a major systemic change programme in Asia with a view to assessing ‘what works’. The conclusions of the evaluation were highly critical of its impact. However, the evaluation made a number of fundamental errors in the way the evaluation was conducted which led them to draw inaccurate conclusions.

Firstly, the evaluation was assessing the programme based on criteria which were inappropriate for systemic change programmes. Instead of looking for how the system had changed and the impact this was having on poverty, evaluators examined the number of direct beneficiaries of programme activities. Ideally, a systemic change programme would not be visible to those that benefit from it in the long term, working through partners to catalyse change.

Secondly, the programme was criticised for not fully addressing the issues of producers in specific sectors. However, the programme’s focus had moved on as it began to concentrate on different sectors with pro-poor growth potential and interlinked markets where systemic constraints were present.

Thirdly, the evaluation was conducted soon after interventions had tended, significantly underestimating impact. Systemic change programmes result in incremental improvement with a significant lag between the change implemented and their effects being realised and so the timing of the evaluation was never likely to capture this. There was, therefore, a misunderstanding in the way the evaluation was conducted in relation to the programme’s objectives. This questions the parameters of ‘independent’ evaluations particularly in the context of fluid, complex and dynamic programmes (Interview with key informant involved with the programme).

4.1.4 What’s the Problem?

The contested subject of malarial bednets is one where the incompatibilities of evidence-based policy and systemic change are clear. Whilst this paper will not engage with the case study in detail as it is documented extensively elsewhere, it is useful to note the dynamics of the debate and its relevance to this paper. A number of studies have used RCTs to examine these effects. These studies found all of the assumptions made about paid distribution models (recipients valuing the nets more highly, reduction in resale etc.) grossly overstated, if not entirely false (Dupas, 2009; Hoffmann et al.,
Conversely, researchers including, notably, Prof. Christian Lengeler (Abdulla et al., 2001; Lengeler et al., 2007; Heierli and Lengeler, 2008) examine the effects of a wide range of intervention models across a number of countries finding pluralistic results in the models of ‘what works’ and what does not. In Tanzania, for example, a private sector model has proven highly successful and sustainable for different at-risk groups and, additionally, has provided increased incomes for those involved in the manufacture of the nets.

The requirement to distil programmes to binary quantitative assessments means that, in the context of a short term programme with short-term goals, free distribution of nets results in a greater reduction in rates of malaria than a programme which aims to build local capacity to cater to this demand and funding is allocated accordingly. This evidently has consequences for the sustainability and scale of change that can be brought about. This is a clear example of the point referred to earlier in this section where there is no shared agreement of the problem (Harrison, 2000). If the problem is defined as being the incidence of malaria in a given area within a given timeframe then there are definitive ways to evaluate approaches to that. If the problem is defined as the inability of a local economy and health system to cater to its health needs in the long term – including the long term incidence of malaria - then a different approach is required together with different evaluation techniques.

Evidence-based policy is, then, striving for the wrong goal; results are serving as a disciplinary tool which compromises the capacity for innovation amongst programme designers and implementers, while evidence is striving for a solution to development challenges. Instead results could be used in tandem with evidence to assess the efficacy of different approaches to determining and tackling context specific problems, in line with broader development objectives.

5. Positive Signs and a Way Ahead

To this point the paper has documented the rise of evidence-based policy and the results agenda and the impact of this on development programming. Subsequently the paper identified a welcome move towards longer terms thinking and systemic change objectives in development programming but highlighted the incongruity of these two trends. Ultimately, this demonstrates the cognitive dissonance within the minds of policy makers.
This could lead to a nihilistic outlook on measurement within development programmes in general but when compared with the alternative of wasted money, ineffective programmes and poor results it is clear that an alternative must be sought. That something is difficult to measure should be motivation to find an alternative approach rather than to abandon it all together. Ways of addressing this problem include a reappraisal of qualitative methods, triangulation of quantitative and qualitative methodologies and the incorporation of new and emerging techniques which aim to synthesise the two. While it is beyond the scope of the review to engage in thorough methodological discussions, participatory statistics and a range of other ways of quantifying qualitative data offer new and robust measurement techniques which take greater account of context.

Furthermore, the limitations, and unintended impacts, of the results agenda are now beginning to be recognised by some donors. USAID have, in recent years, begun to advocate the use of the ‘Degrees of Evidence Framework’, in agricultural value chains and finance programmes. Based on programme experience, evaluators began to recognise the need for flexibility in the application of methodologies to different social, economic and cultural contexts. The framework provides those working on programmes with a means of appraising the best practical way to obtain robust data within the context and resource constraints of any given situation. This represents progress towards a better understanding of how to evaluate change. The advocacy of triangulation and pragmatism in the assessment of methods too is positive. However, the hierarchical assessment of methods within the framework are not appropriate to the complex unbounded problems of the majority of systemic change programmes and its emphasis on targets means that it still fails to capture many of the externalities which are in fact one of the primary objectives of systemic change programmes (Creevey et al., 2010).

The DCED standard has also been useful in guiding development programmes towards better measurement of their impact. From a place where measurement of and comparison between development programmes was almost entirely absent, the standard sought to establish a minimum level for the measurement of programme effects. However, the emphasis in the standard on the need for measurement to be an internal function runs contrary to the drive amongst many donors for independent evidence and increasingly, particularly amongst some donors, the minimum is not good enough. Furthermore, the standard is indicative rather than instructive and, as a private sector development focused initiative, there is very little indication of how to fully capture change at a system wide level.

More widely, evaluations are, in some cases, encouraged as iterative, collaborative and adaptive processes developed between evaluators and implementing parties. In an even
smaller number of cases, new and innovative approaches to the measurement of complex development programmes are being introduced and, more importantly, viewed as acceptable by donors. This could lead to the development of a more pragmatic and less positivist best practice standard for results measurement in systemic change programmes, which can begin to contribute to the ‘evidence’ on which policy is increasingly based. However, such evaluation designs remain rare and rely on courageous individuals within donors to advocate for their employment in the face of normative voices within their organisation.

In order to reconcile the clear incompatibilities between how evidence-based policy and the results agenda are currently realised in development with the parallel trend toward a systemic change approach, there is a clear need for further research, conceptual and methodological development of how to capture change without compromising programme objectives. The majority of development programmes are open to measurement to be used as tool in the improvement of development outcomes. However, this should not occur at the expense of innovation and pursuing optimal developmental outcomes. Evidence needs to be redefined as a means of assessing approaches to development so that lessons can be learned and adapted to different contexts rather than a means of assessing which tool should be applied universally.

References


---

8 Advocates of a presumed and unverifiable orthodoxy i.e. those that believe that quantitative, experimental/quasi-experimental, ‘independent evaluation is the only defensible means of assessing efficacy.


DAC 2010. OECD DAC Peer Review of the United Kingdom. In: DEVELOPMENT ASSISTANCE COMMITTEE (ed.).


DFID 2011. How to Note: Reviewing and Scoring Projects. In: DEPARTMENT FOR INTERNATIONAL DEVELOPMENT (ed.).


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, Penguin.


KARLAN, D., GOLDBERG, N. & COPESTAKE, J. 2009a. 'Randomized control trials are the best way to measure impact of microfinance programmes and improve microfinance product designs.' Enterprise Development and Microfinance, 20, 167-176.


ROGERSON, A. 2011. What if development aid were truly “catalytic”? *ODI Background Note*.


SOLESBURY, W. 2001. Evidence based policy: Whence it came and where it’s going. ESRC UK Centre for Evidence Based Policy and Practice London.


