Development Policy Review, 2016, 34 (1): 5–27

Probing for Proof, Plausibility, Principle and Possibility: A New Approach to Assessing Evidence in a Systematic Evidence Review

Anouk S. Rigterink and Mareike Schomerus*

This article proposes a new approach to assessing evidence during a systematic evidence review aiming to inform international development policy. Drawing lessons from a number of social science systematic evidence reviews, the article identifies how the method's limiting perspective on evidence (including the exclusive focus on 'gold standard' empirical information) has serious disadvantages for the usability of evidence reviews for policy. This article aims to provide an alternative framework that allows for a less exclusionary, yet policy-practical, way of assessing evidence. We propose four perspectives on evidence, appropriate for different stages in the policy process: principle when setting or prioritising broad policy goals, plausibility when assessing specific future policies, proof when evaluating past policies and possibility when striving for innovation and allowing exchange of ideas.

Key words: policy, systematic evidence review, proof, plausibility, principle, possibility

1 Introduction

Does international development policy want to play it safe? An emerging trend towards basing policy decisions solely on evidence of what works gives that impression. A renewed emphasis on measuring success and failure in policy and development programming, as well as a drive to make development aid cost-efficient, has elevated 'evidence' to a development policy kingmaker. Donors assume that basing policy decisions on evidence 'increases the success of policies, their value for money and their impact' (Gasteen, 2010: 1).

The pursuit of 'evidence' in policy-making has also shown that there are no clearly signposted roads for a policy-maker to follow when hunting down 'evidence'. Indeed, often it is not clear what 'evidence' looks like, so the search warrant has no description or picture of what needs to be found. In other words, although vast

*Respectively, Research Officer (a.s.rigterink@lse.ac.uk); and a Research Fellow, London School of Economics and Political Science, Houghton Street, London, WC2A 2AE (ms@mareike.net). The authors are grateful to Hakan Seckinelgin, Henry Radice and the anonymous reviewer for the helpful feedback, to our colleagues at the Justice and Security Research Programme with whom we conducted the evidence reviews and to those students at the LSE Department of International Development who participated in seven systematic evidence reviews. This research was funded by the UK Department for International Development.

amounts of information and knowledge exist about development programme approaches that fail or succeed, including whether failure and success is the same for those implementing a policy and those at its receiving end – such knowledge is often not systematically accessible, usable or appropriate.

Evidence-based policies require two things that are surprisingly difficult to come by: solid and appropriate information alongside solid systems able to administer, assess and disseminate such information. Our experience conducting a number of systematic evidence reviews in international development – which forms the backbone of this article – has taught us that these solid systems are not in place. Researchers, policy-makers and practitioners will, even if information is easily accessible, inevitably return to the question of what kind of information constitutes 'evidence' and how to interpret this 'evidence' for a policy. Systematic 'evidence' reviewing is one answer to this question. We learned that this solution has a number of disadvantages: by defining 'evidence' as exclusively empirical information, theoretical knowledge and often seminal work is excluded while innovation is potentially inhibited. More importantly, it is not clear how one moves from finding 'evidence' in a systematic review to making this usable for policy decisions. What is needed is a way of clearing the pathway of moving from reviewing the state of the evidence to assessing its contribution in a less exclusionary way that allows a hands-on debate regarding the role particular evidence might play in the formulation of policies. What we propose in this article is a less exclusionary framework for this process, one that is simple and practical without dismissing how complicated it is to agree on what constitutes evidence. The framework encourages scepticism, but in a constructive way. We see this scepticism as crucial. After all, the emphasis on evidence stems from the belief that 'clean, clear information' will be the main driver of change in the future (White and Waddington, 2012: 351). One problem, however, is that in systematic evidence reviews the evaluation criteria are exclusively focused on assessing the quality of evidence based on a very limiting set of quality criteria.

With this article, we propose a different approach to assessing evidence in systematic evidence reviews: rather than looking at its *quality* through often dubious criteria, we propose to use systematic evidence reviews to assess evidence according to its *usability*.

Yet the quest for this information obscures the fact that the word 'evidence' is an umbrella term that describes a broad category; only with further specification does the notion become meaningful within the process of systematically reviewing evidence. Speaking about 'evidence' in this context is like speaking about 'countries' whereas in fact it makes a big difference if you are talking about Bhutan, Brazil or Scotland. Thus rather than use the word 'evidence' to pretend it describes something specific, we propose four distinct interpretations of evidence when systematically assessing it: proof, plausibility, principle and possibility. The second part of the article describes in detail what we mean by these. Each of these sub-categories of evidence warrants inclusion in a systematic evidence review, yet when it comes to translating the findings from an evidence review into policy thinking each sub-category is appropriate for a different policy process. Having established this new vocabulary, we are going to discard the inverted commas around 'evidence' as the

word now serves only to describe a broader phenomenon, rather than being implicitly imbued with different meanings.

In providing a framework for a less exclusionary approach to assessing evidence in a systematic review and for clarifying its applicability for specific policy purposes – like moving from talking about countries to talking about Bhutan or Brazil – we pay special attention to how and when evidence falling short of the 'gold standard' could be used in policy-making. This gold standard, referenced by the UK Department for International Development (DFID), refers to experimental research that seeks 'to demonstrate cause and effect relationships ... because they construct a "counterfactual", experimental studies significantly reduce the risks of important biases affecting the findings of research' (Department for International Development, 2014: 7). Since this gold standard is not always achievable – nor its categorisation as gold unchallenged – we aim to identify ways of recognising situations where the drive for evidence-based policy is unproductive. The remainder of this article will thus tackle a few questions: what are the shortcomings of systematic evidence reviews in the social sciences? When systematically reviewing evidence in international development, what information constitutes what kind of evidence? How can this evidence be assessed in a less exclusionary way when probed for its policy relevance?

1.1 In a systematic evidence review, what is evidence?

Systematic evidence reviews are a relatively new phenomenon in international development, maybe also because, as White and Waddington (2012) argue, 'the evidence bar has been rather low, with many policies based on anecdote and "cherry picking" of favourable cases' (ibid.: 351). Systematic evidence reviews are meant to counter cherry picking by investigating all the information that is out there in a uniform way. This is meant to separate poor quality evidence from good quality, akin to an archaeological dig during which lots of stuff is unearthed, but the decision as to what constitutes a valuable item worth preserving as opposed to another to be discarded is made with a specialist set of criteria. Systematic evidence reviews are after empirical evidence of a certain kind: DFID lists 'primary experimental', 'quasi-experimental' and 'observational' as acceptable types of evidence (Department for International Development, 2014: 9).

This is a narrow definition of evidence, and one that is problematic in itself. Furthermore, it overlooks the fact that not all empirical observations are equally convincing, even to those who operate within narrow definitions of evidence. Various scales exist to demarcate the information hierarchy, delineating where a particular piece of information stands within this hierarchy. The peak of the hierarchy is the aforementioned 'gold standard' referring, in DFID's definition, to research designs that 'explicitly seek to demonstrate cause and effect relationships, and are able to do so with varying degrees of confidence. Because they construct a "counterfactual", experimental studies significantly reduce the risks of important biases affecting the findings of research, and for this reason, they are often regarded as the "gold standard" for research which aims to isolate cause and effect.' (ibid.: 7)

The Maryland Scale of Scientific Methods is one of the best-known tools for assessing this information hierarchy. To score three or higher on this five-point scale, a study must have a treatment and control group and must record outcomes in each group before and after execution of the policy under study. Some authorities regard this as the minimum standard for drawing conclusions about the success of a policy. A study also including relevant control variables would get a score of four, whereas the highest score is reserved for studies in which the policy is randomly assigned between treatment and control group (Farrington et al., 2002). This study design, a Randomised Controlled Trial (RCT), is also considered the 'gold standard' in research design (Castillo and Wagner, 2013). However, it is evident that this standard is best suited to a particular type of research: quantitative research with policies that can be (ethically) randomised. Policies at the individual or community level, such as medical interventions (Miguel and Kremer, 2004) or interventions in schools (Paul et al., 2009) can be relatively easily randomised and have been the subjects of RCTs. Policies at a higher level, for example the global Kimberley Process Certification Scheme to counter so-called conflict diamonds, cannot be randomised easily and control groups may not exist. Consequently, evaluations of these policies have used methods that would score below three on the Maryland Scale (Grant and Taylor, 2007). Finally, in order to place a study on the Maryland Scale, the study has to be transparent about the research methods used. As will be shown in the next section, numerous studies are not.

This perspective on evidence also overlooks some of the realities of evidence use. Despite gold standards, in the heat of searching for evidence to include in policy documentation the origin of information and how it was collected is often neglected. When it is convenient, someone else having said something supportive of a policy approach can quickly become evidence in the echo chamber – even if no systematic evidence review convincingly brought up evidence. Or, as Andreas and Greenhill (2010) illustrate, objective-sounding statistics can be the result of a process in which erroneous numbers are repeated so often that they become understood as fact. Distinguishing between the noise echoing off the sides of the consensus mountains and empirical observations is not always easy.

2 Lessons from systematically reviewing evidence

Our starting point for thinking about these questions was a systematic evidence review – a mechanical assessment of information. The main donor of our research programme, the UK Department for International Development (DFID), required such a review, although it did not impose a specific method. The systematic evidence review was meant as a first step in designing the research agenda: reviewing all available evidence would expose areas where evidence was scant and gaps needed to be filled. The research programme conducted systematic evidence reviews on seven topics in the field of justice and security in conflict-affected areas using broadly the same basic method (Carayannis et al., 2014; Cuvelier et al., 2013; Forsyth and Schomerus, 2013; Luckham and Kirk, 2012; Macdonald, 2013; Seckinelgin and Klot, 2014).

Detailed descriptions of methods, including how they varied, can be found in the individual papers. We largely followed general recommendations to pursue 'a

well-defined question for the review, (2) an explicit search strategy, (3) clear criteria for the inclusion or exclusion of studies, (4) systematic coding and critical appraisal of included studies, and (5) a systematic synthesis of study findings, including meta-analysis where (it is) appropriate' (White and Waddington, 2012: 354). Each systematic review consisted of two steps. First, we systematically searched literature on the particular topic that met our criteria. Second, we systematically assessed the quality of all the works in this pool of literature. Implicitly, this involved making judgments on which information makes for better evidence. The outputs from this process were reports on the state of the evidence base in the seven selected areas.

Online search tools and databases were our major source of information in the first step. These mainly included search engines geared towards searching academic literature, but also a few specialising in grey literature – that is literature that is not peer-reviewed and is generally self-published by the organisation conducting the research. We entered an elaborate and standardised keyword string in each search engine. Search results were then filtered manually. Any studies containing empirical information at the individual level were added to the pool of literature, any studies not providing empirical information were discarded.

In assessing the quality of research works in the pool, we used a number of criteria to judge the quality of the evidence presented in the studies found. These criteria were formulated differently for different studies employing different methods. Using this method we assessed more than 1000 articles for all seven systematic reviews combined.

The importance of systematic evidence reviews for social sciences is steadily growing, indeed they have been called 'one of the great ideas of modern thought' (White and Waddington, 2012: 351). It is generally assumed that reviewing information systematically 'strengthen[s] the capacity to make evidence-informed decisions by providing rigorous and timely assessments of the evidence base to decision makers' and to 'make it easier for policy makers and practitioners to develop evidence informed policy' (Department for International Development, 2013). Yet there are many caveats to this assumed chain of events; a few of these caveats relate to the limitations to systematic evidence reviewing. These limitations might be specific to the topics we explored, but they could plausibly hold for the method in general as others have shown (Mallett et al., 2012).

2.1 Lesson 1: Systematic evidence reviewing runs the risk of omitting seminal works

We had expected that designing a systematic way to assess the quality of works of research would be difficult – it is difficult to agree on quality criteria with team members from different academic disciplines. We treated the first step of constructing the literature pool as a largely technical exercise: make keywords, put them in search engines, sift through the results.

It turned out that the most difficult element, with the strongest influence on the outcomes, was not assessing the quality of the studies, but finding them –

which is important since including 'all the available studies, including journals, grey literature and unpublished studies' (Gasteen, 2010: 1) is considered the cornerstone of systematic reviewing. Works that we considered seminal did not end up in the pool, even after using more than ten search engines and carefully constructed keywords. Frustratingly – and tellingly – a lot of our own work did not show up through the searches either. Colleagues whom we asked for feedback were often mystified by the absence of what they considered obvious references. In addition, we found very little relevant grey literature. Search engines alone were clearly no guarantee of finding 'all available studies'. One group that used 'snowball searches' and a separate inquiry amongst peers succeeded better at capturing all relevant literature; but at the cost of making the process less systematic. Further, by focusing only on publications a publication bias sets in: since 'null, and possibly negative, findings are less likely to be published' (Waddington et al., 2012: 362), we are less likely to ever hear about what could potentially be crucial findings.

Why did the systematic review process miss so much? Maybe not all authors provided accurate and sufficient keywords. There are no universal repositories for working papers, 'grey' literature and papers that have yet to go through the time-consuming process of being published. It has been recognised that cataloguing research in international development is 'less standardised amongst a wide array of role-players, with no equivalent of the large freely accessible health care library, PubMed.' (Stewart et al., 2012: 433-4) Further, in our systematic review process we only included papers with individual-level empirical information that some papers, however seminal, may not contain.

An even more likely reason is the fact that it is extremely difficult, if not impossible, to capture certain fields of research in a set of keywords. This could be because different disciplines indicate the same subject with different keywords. For example, large amounts of literature exist on peacekeeping – indeed there is a journal primarily devoted to the topic[1] – but within economics or quantitatively-oriented political science, one rarely finds the term. A leading review of the state of the literature in these fields (Blattman and Miquel, 2010) tellingly only mentions the term 'peacekeeping' when it discusses efforts to integrate historical-political case studies and statistical methods.

Fields of study, and thus keywords, may change over time. Truly seminal works may create entirely new fields of study, which are only later captured by keywords. Insights from these works may be taken beyond the context in which they were originally proposed. Charles Tilly for example is now often cited in studies on conflicts in Africa for his theory that war can build states, which was published under the title *Coercion, Capital, and European States: AD 990–1992*. Its abstract does not mention the keywords 'war', 'conflict' or 'violence'.

1. Journal of International Peacekeeping

2.2 Lesson 2: Restricting the definition of 'evidence' to 'empirical information' is limiting

Many shortcomings of systematic evidence reviews are a consequence of defining evidence as exclusively empirical information, which is how DFID implicitly, and sometimes explicitly, defines it when speaking about making decisions based on 'what we know' and how this increases 'the success of policies, their value for money and their impact' (Department for International Development, 2013). In the past, a prominent criticism of systematic evidence reviews has been their narrow focus on quantitative evidence (Snilstveit, 2012). While we specifically moved beyond that, we found that the criterion that a study had to contain empirical information at the individual level might be equally restrictive: it generally reduced significantly the number of results from the search engines. The evidence paper on conflict resolution, for example, reports that of the 543 studies found by one search engine, only 63 (12%) were considered to include individual-level empirical information from reading the abstract. For another search engine, this was 60 out of 1635 (4%), for another 56 out of 1060 (5%). Typically, an additional third of the papers initially included because of their abstract did not include a meaningful amount of empirical information upon closer reading.

Of course, our self-imposed inclusion criteria required studies to have individual-level empirical information, not empirical information in general. However, from our experience with the review process, a large percentage of publications in academia are non-empirical – or as Waddington argues, there is 'a paucity of primary studies which are able credibly to address causality' (Waddington et al., 2012: 380) – and the number of empirical works is so small that this criterion ended up being interpreted very widely. Non-empirical works include purely theory-based work, opinion pieces and studies that are unclear about the information they use. Although defining only empirical research as evidence would appear sensible, discarding theoretical works for the purpose of a systematic evidence review must mean that some babies get thrown out with the bathwater.

The following realisation is important: if evidence is strictly defined as empirical information, and all policies have to be evidence-based, there can never be a new type of policy, strictly speaking. Empirical information by definition requires some real-world phenomenon to be studied, and if the policy has never existed it cannot have been. Although lines may not be so clearly drawn in reality, it is a realistic concern that the focus on evidence-based policy may restrict innovation.

Theoretical work may be useful for policy-making, especially when few empirical studies exist. If empirical evidence is not available, a policy that at least could work according to a sound theoretical framework would be preferable to one that is not supported, or is actively contradicted, by theory. Unless we are willing to accept that no policy is possible in the absence of empirical observations, then theory or principles can be relevant. The framework presented in this article aims to

bring theory, and innovative or principled thought, back into the so-called evidence-based policy process.

2.3 Lesson 3: Studies (especially in certain disciplines) often do not contain enough information on method to assess their quality

Many of the criteria for assessing the quality of a study rely on the author describing his or her methods in some detail. However, we found that many studies, especially from the more qualitative disciplines, neglected to give the reader even basic methodological information – although Duvendack et al. (2012) report similar concerns about lack of information on design and analysis in quantitative work. In some cases information was presented as fact, or as stemming from the author's research, without further explanation as to how this information was obtained and filtered by the author. It was also common for studies to mention in passing that they were based 'on interviews', but not on how many interviews, with whom or where these were held and by whom. A number of studies that we knew to be based on sound methods (either from knowing the author or seeing him or her presenting the work in person) received very low scores on the quality assessments because they did not make the method transparent. From a policy-maker point of view, although it may be difficult to judge the quality of certain studies, this does not mean that these studies need to be automatically dismissed as non-evidence. Especially when studies about a particular topic are rare, it would appear to be a waste to omit this information from the policy process. Yet if researchers wish their work to be classified as high-quality evidence, they need to describe their methods and any shortcomings in detail.

2.4 Lesson 4: The missing step: an assessment of the overall state of the 'evidence'

Systematic evidence reviewing does not necessarily provide a clear analytical path for moving from reviewing numerous individual studies to a general assessment of the state of the evidence (Hannes and Macaitis, 2012). We designed no systematic guidance on how this should be done, thus the process of moving from finding all relevant information and assessing its quality to assessing the quality of the evidence base as a whole remained a black box – a shortcoming in transparency that has also been noted by others (Snilstveit et al., 2012). Within the team, some authors had a tendency to disengage from the systematic legwork that had been done and in their writing drew on their own knowledge. This experience is not unique to us. 'Synthesis' or the inability to do meta-analysis and draw meaningful conclusions from an evidence review – without being guided by pre-existing knowledge or opinions – and difficulties in '[g]enerating useful policy recommendations' are two of the seven practical challenges of the systematic evidence review method mentioned in another review on systematic evidence reviews (ODI Secure Livelihoods Research Consortium, 2012).

It is not the case that methods to synthesise results from different studies do not exist. However, these methods can only be applied to a very restricted set of systematic reviews. Meta-analysis is one such method to synthesise results from quantitative studies. It involves running a regression on the regression results obtained by the studies under review (Mekasha and Tarp, 2013). As such, it can only be applied if there is a substantial body of studies with the same outcome variable, the same explanatory variable and the same method. It is most frequently used in the medical field, where numerous studies investigate a single treatment for a single condition. Its applications to the field of development, and especially to the topic of insecurity and injustice, are limited: even if it is possible to agree on the 'condition' to be treated (poverty, insecurity, lack of justice), potential 'treatments' are myriad and fast-changing, and comparable studies scarce.

A second approach to synthesising is 'narrative synthesis'. While no consensus exists as to what exactly a narrative synthesis is, this approach broadly summarises a given topic from many angles – thus giving a comprehensive narrative of the state of research – but without consideration of a statistical or systematically comprehensive summary of the state of evidence. Narrative syntheses tend not strictly to follow a methodical search for evidence; this means they are often not transparent about how decisions about relevance and strength of evidence were made (Collins, 2004: 102). However, more systematic variations of this approach can now also be found (for example (Greenhalgh et al., 2005; Hannes and Macaitis, 2012)). Furthermore, narrative synthesis has been found to be possibly most effective when used in conjunction with meta-analysis (Rodgers et al., 2009). Below we outline a particular case in which narrative synthesis might be the best way of presenting a review of the state of knowledge on a topic.

3 Thinking about systematically reviewing evidence for policy relevance in a different way

Apart from learning more about the need for transparency, clever choice of keywords, the shortcomings of search engines and how difficult it is to write neutrally based on information presented, rather than in an opinionated way, we learned another crucial lesson. It is actually not clear how a systematic evidence review would automatically translate into a simple answer to what the state of the evidence implies for a particular policy. Although we made lots of interesting points in a substantial number of evidence reviews, we have so far not had policy-makers banging on our doors for more. Maybe that is because what we presented was just enough to draw on for some decisions – after all, it has been noted that policy-makers need to make decisions, even if the evidence base on which to make those is limited (Snilstveit, 2012: 391). Or maybe we did not come up with a hands-on practical framework for how the kind of evidence we found might be relevant for policy thinking. Maybe we felt as if we were asked to render complex findings into information popcorn – fluffed up, making a bang and easy to eat. That, however, does not get us away from the need to provide a better understanding of the information that is out there. We concluded that doing systematic evidence reviews

becomes a valuable exercise when used to think about the evidence we find in different ways. Sifting through large amounts of research just to see if individual level empirical material exists, or if RCTs have been conducted, misses a broader point. A better way of treating systematic evidence reviews is to engage with the evidence found with the aim of gaining a better understanding of how a particular issue is researched, understood and thought about. Rather than being like popcorn, this is more like the laborious, messy, but satisfying, preparation and consumption of corn-on-the-cob.

We understand why this is rarely done. The corn-on-the-cob approach is missing because of the different claims researchers and policy-makers aspire to make. In the words of Nancy Cartwright: researchers want to make 'it works somewhere' claims, studying individual cases using methods geared to making sure that conclusions are valid for the case studied (internal validity) (Cartwright and Munro, 2010). Policy-makers, on the other hand, would prefer 'it will work for us' claims on which to base their decisions (ODI, 2009: 2).

These two modes of working seem irreconcilable – and yet we are proposing a method of translation to bridge this gap when conducting systematic evidence reviews. It starts from the realisation that the evidence debate among those in research, policy and practice could benefit from agreeing on certain terminology to describe what particular pieces of information are and are not doing. We are making a controversial proposal for the use of a research-sympathetic and policy-friendly framing of how to think about good evidence when faced with huge amounts of research on topic. Thinking practically about evidence in the way we propose here might allow for better categorisation of information, more honesty when dealing with information and, with that, hopefully more transparency and better decision-making.

Our suggestion is to systematically review evidence and assess its applicability for informing policy by dividing it into 'The Four Ps': proof, plausibility, principle and possibility. Each of the Four Ps, we argue, is relevant in a different stage of policy thinking, thus systematic evidence reviews would be more useful if they divided their findings according to these categories:

- When setting or prioritising broad policy goals, we argue that an emphasis on searching for evidence as it is currently defined in systematic reviews is misplaced and propose *principle*.
- When assessing different possible future policies to attain a specific broad goal, *plausibility* is the most relevant category in synthesising existing evidence.
- We see limited use for the notion of evidence as *proof* outside the evaluation of past policy – the idea that systematic evidence reviews will propose proof on a matter is unhelpful.
- And when striving for innovation, we propose *possibility*, which is a category of evidence that tends to be discarded in systematic reviewing.

In all these stages, we emphasise transparency about *what* a particular piece of information constitutes evidence *for*, and with what level of certainty.

3.1 The preferred P: proof

The UK's Minister for International Development at the time of writing, Justine Greening, laid out her vision for her department's future in her first major speech after her appointment. For her, international development supported by the UK entails UK aid money being spent 'in the right places, on the right things and done in the right way' (Greening, 2013). Establishing a clear distinction between something that is right and wrong implies foolproof knowledge as to what makes a plan, a time and a place 'right'. We name this approach to evidence *proof:* the notion that one can search for evidence that, once used, will guarantee future success. The notion of *proof* before execution of a policy is unhelpful, not just because we learned from doing systematic reviews that the required information is difficult to come by or of low quality. It also assumes that policy information that can act as *proof* could possibly exist.

This interpretation of evidence is reminiscent of that used in a criminal court, where a piece of evidence is presented and, once a judgement has been made on it, the evidence has become proof for the correctness of that judgement. How problematic the notion of proof is in situations when making decisions on policies to be implemented in the future is obvious. What proof-based policy is trying to do is to predict the future: because something is considered to have been proven in the past, it is expected that the research or practice community will deliver continued proof that it will work in the future. Yet, in the court of proof-based policy, the evidence stems from studying a different incarnation of this policy, at a different time, often in a different country. '"Cutting out the noise" probably misses the point in international development research (and the social sciences more broadly), where context is everything', is how researchers at the Overseas Development Institute (ODI) have described this issue (ODI Secure Livelihoods Research Consortium, 2012). If we can find a study including convincing evidence on Sri Lanka, can this be generalised to other contexts? A policy's success in Sri Lanka does not guarantee success in Sierra Leone, nor even continued success in Sri Lanka itself. To stick to the court analogy: proof in one legal case, even gold standard proof, cannot be called upon to make the same point in another case with different actors, times and places. The notion of exchangeability contradicts everything we know about context awareness. Crucially, from a research perspective, proof implies that a final verdict on a matter has been reached, rendering future research and learning – and maybe even development – unnecessary. But there is no need to worry just yet: numerous systematic reviews led to the realisation that the search for proof-based evidence for future development policies mostly draws a blank.

The notion of evidence as proof crowds out flexibility and responsiveness. These two aspects are crucial in development and difficult to underpin with information. Finding proof for something that will work in the future directly contradicts notions of empirical research, which requires observation and experience. Using the notion of evidence as proof when deciding on future policy is a true head scratcher. Seeking evidence that can be used as *proof* for future success is not only a

tilting at windmills, but a confusing admission that windmills are in fact a real enemy.

We do see scope for evidence as *proof* when it comes to evaluating past policy; and this would be a useful way of reframing the task of a systematic evidence review. Empirical observation is possible with past policies; attempting to get as close as possible to the gold standard of *proof* might be useful in this context. However, when attempting to create *proof* it should be clear what particular information is proof *of*. Information can only be considered proof in an extremely limited context. Even in gold standard statistical research, proof is only provided within a certain confidence interval. Observing a relationship between policy and outcome at the 5% confidence level loosely means that there is a probability of 5% that this relationship is due to chance. A randomised controlled trial can only provide *proof* applicable to the past policy studied, executed in the same way, at the same time and in the same place. It can only make an 'it works somewhere' claim (Cartwright and Munro, 2010). Sometimes, the very way in which a policy is executed is affected by randomisation.

Consider, for example, a study investigating whether election campaign messages based on clientelism or on national public goods are more effective in winning votes in Benin (Wantchekon, 2003). This involved convincing politicians to change their campaign messages in randomly chosen election districts to include exclusively clientelistic or public goods arguments. What was being studied here was the effect of enforcing particular campaign messages in random districts in a particular election circle in Benin. For this, the study provided evidence in the sense of *proof*. What the study does not provide proof of is the impact of clientelism (as a strategy chosen by politicians based on whether they think it is likely to win them votes) on voting behaviour. Again, we do not mean that results of Wantchekon's study are meaningless when attempting to answer the latter question, but, at best, it can make the existence of such a relationship more *plausible*.

3.2 The realistic P: plausibility

Plausibility is an alternative way of navigating the translation process from systematic evidence reviewing to policy relevance of the information found. Plausibility allows for consideration of theory, or empirical studies that are of reasonably high quality yet fall short of the gold standard as evidence. Plausibility is not proof, but is broader than most current definitions of evidence. An argument that a policy can plausibly work may include retrospective proof from other contexts (proof that it was a right thing, done in a right way, at some other time and/or in some other place). But plausibility goes beyond this: it has to demonstrate that it is probable that the policy can work in another time and at another place. This is akin to what Cartwright terms a capacity claim: the claim that a treatment has a relatively stable capacity to provide a certain outcome (Cartwright and Bradburn, 2011; Cartwright and Munro, 2010). Plausibility comes with a health warning: it still requires the huge effort involved in conducting a systematic evidence review, but in translating this to policy-relevant information it delivers no

guarantees, no one-stop evidence shop, no generalisations. It means more scanning of the horizon to continuously gather information to see if what seemed plausible at first really holds up – so the systematic evidence review is nothing but a snapshot in time. If policy-makers and practitioners invoked plausibility when considering future policy, research could more credibly feed into programme planning, designing plausible Theories of Change, and allowing for context-specific programming. Any how-to guide on using evidence for policy and practice would be more convincing if it stressed the importance of plausibility; any systematic evidence review would be more complete if it synthesised from the perspective of plausibility.

An obvious way to start building a plausibility case is to specify which activities the policy or intervention will entail, and what it intends to accomplish. To fully assess the intervention's plausibility of success, more is required than a pairing of activities and outcomes. A good plausibility case includes all the steps that link the activity and the outcome to each other. These together form a theory of how the policy is supposed to work, factors that help and hinder the effectiveness of the policy and the interaction of the policy with other factors. This is central to establishing capacity (Cartwright and Bradburn, 2011; Cartwright and Munro, 2010). Systematic evidence reviewing may detract attention from constructing such a theory, as it might misplace the focus. In the words of ODI:

> Outcomes are ultimately shaped by programme design and delivery, as well as by context [...] and SRs [Systematic Reviews] do not necessarily help us understand these dimensions. In other words, the question of why things work is just as policy relevant as whether or not they do in the first place. (ODI Secure Livelihoods Research Consortium, 2012)

Consider for example, a food for education programme (FEE). A plausibility case would state that giving households food in return for school enrolment increases educational attainment, and then it would detail exactly *how* this is expected to work. It would not only state that giving children meals at school can improve their learning, but painstakingly set out why that is the case. We may expect FEE programmes to improve children's nutritional status, and better-fed children to have increased cognitive ability. Alternatively, poor households may not be able to invest as much in education as can be considered optimal in the long run and FEE programmes may relieve this budget constraint, increasing enrolment and educational attainment. It is also possible that a plausibility case brings to light mechanisms through which FEE programmes may *decrease* educational attainment: increased school enrolment may lead to overcrowded classrooms and decreased learning, or increased school enrolment may reduce child labour, decreasing household income and thereby the nutritional status of children along with their cognitive ability (Adelman et al., 2008).

This is not a new suggestion; a number of donors now ask aspiring projects to formulate a Theory of Change, which includes such elaborate descriptions of the steps between activities and impact (Stein and Valters, 2012). Detailing these steps can help to find material to build a case for or against plausibility when doing a systematic evidence review. There may be very few studies done on the full

intervention proposed. For example, Adelman et al. conclude that in 2008 gold standard studies into the impact of FEE programmes on educational outcomes were 'relatively few' (Adelman et al., 2008). However, numerous studies may exist on the links in the chain connecting the intervention's activities to an outcome. In our example, Adelman et al. mention: 'nutrition literature offers many more experimental studies on nutrition outcomes than are yet available in the economics literature on education outcomes' (Adelman et al., 2008). Other helpful information in this case might include studies on the impact of nutritional status on cognitive abilities in other contexts, the actual and perceived returns to education in the region of intervention, or children's nutritional status in this region. Developing a Theory of Change and investigating information which may or may not support each of the steps will not prove that the proposed policy will be successful. Yet finding information that supports at least some of the proposed mechanisms through which a policy is thought to work, and that can reasonably be applied to the proposed context, can help make the case that a particular policy could *plausibly* work. Alternatively it can decrease confidence that a particular policy is appropriate.

The act of specifying the mechanisms through which an intervention is meant to work may preclude the need to build a full plausibility case. Even without reference to outside material, some interventions may not pass a test of basic logic. Take a project that proposes to educate women about their rights with the aim of decreasing domestic violence. Thinking this through, this project's success depends on women being unaware of their right not to be subjected to violence in the home and lacking *only* this information to be able to change their situation. While this may be true in some cases, in many others this may not be plausible from the outset. A framing along the lines of plausibility will make these points a lot more obvious when reviewing the state of the evidence.

A plausibility case can include information that falls short of gold standards outlined in the first section. When we recognise the limited use of *proof*, and consequently accept that even gold standard information cannot prove future policy success, the rationale for favouring studies meeting the gold standard over all others becomes less obvious. Are we willing to admit that gold standard information may contribute to a plausibility case to a greater or lesser degree depending on how applicable it is to the time and place where a policy is proposed? If so, why would we automatically exclude from systematic reviews lower quality empirical studies in the same place, or theories that seem particularly appropriate to the context? In a plausibility case this would not make sense; yet the practice of systematic evidence reviewing frequently calls for such exclusion.

Information that may contribute to building a plausibility case could come from studies investigating similar policies to the one proposed, but in another place and time. These may include studies on the intervention as a whole, or on particular mechanisms that are expected to make the intervention work. To this extent, a plausibility case and a systematic evidence review are not very different. However, finding and quoting these studies is not sufficient – that provides neither proof nor plausibility, it just shows that someone has attended class but not done their homework. From basic logic, for studies from another context or another time period to feature in the plausibility case for a new policy it is necessary to explicitly

consider the ways in which the new context is similar or dissimilar to those of the past. What way of executing the intervention was successful in the past? Can this be replicated in the new context? What was it about the context that was conducive to or hindered the past intervention's success? Are similar favourable or unfavourable conditions found in the new context?

This is a productive middle way between blueprinting interventions from one context to another without consideration and completely dismissing proof from other contexts. It requires recognising that evidence is not about proof, but about identifying as probable that a particular action will lead to a particular outcome. Cartwright (2007) calls this hypothetico-deductive reasoning and contrasts this with deductive reasoning A single success of an intervention half-way around the world would contribute little to a plausibility case, whilst consistent studies indicating that the intervention is successful in a range of contexts, including some that can be considered similar to the context proposed, should be a solid basis for plausibility. A range of studies that show that an intervention has been an abject failure does not prove that it cannot ever work anywhere, but does make it implausible that it will.

Another source that can be useful in building a plausibility case is empirical work that is of reasonable quality, but falls short of usual standards for evidence. Some evidence reviews (Mansuri and Rao, 2003) dismiss all studies that do not include a control and treatment group, and in which the intervention is not randomly administered. We see no reason why a gold standard RCT in a completely different context would automatically deliver plausibility, just as we see no reason why a study that contains no RCT should be automatically sent to the scrap heap. The latter study may still contain information that renders the success of another intervention in the same context more or less plausible. This is especially relevant for interventions that are not easy to study, for example the many interventions that cannot ethically or operationally be randomised.

3.3 The adamant P: principle

One common theme in the evidence debate is the tension between pursuing genuine evidence-based policies and seeking out specific information to create policy-based evidence, evidence for policies that were already decided on for other reasons – a risk of which policy-makers are well aware (House of Commons, 2006). These reasons could include ideological concerns or simply because policy-makers get caught up in their own challenges when attempting to use empirical evidence to decide on polices. Vince Cable, UK Secretary of State for Business, Innovation and Skills from 2010 to 2015, described these challenges as grappling with the need for speed, superficiality, spin, secrecy and, from their own perspective, scientific ignorance with regard to methods and testability (ODI, 2009: 1-2). Phil Davies, former Deputy Director of the Government and Social Research Unit in the UK Cabinet Office, said that policy-makers act on 'their own values, experience, expertise and judgement; the influence of lobbyists and pressure groups; and

pragmatism', creating a situation in which 'researchers and policy-makers have completely different concepts of what constitutes good evidence'. (ibid.: 2)

Thus, we know that policies are not always based on evidence, and sometimes evidence is shaped around policies. However, this does not necessarily mean that the entire debate around evidence is disingenuous. Although the practice of policy-based evidence *could* be a sign of policy-makers' distaste for engaging with evidence, it may simply be a sign of an entirely different logic of decision making – one that can be identified as *principl*e.

It is important to recognise what *proof* and *plausibility* can and cannot do when searching out information through systematic evidence reviews. Proof is useful to evaluate whether a past policy has obtained its goals. Plausibility details whether it is likely that a proposed policy will obtain its goals. However, neither tells us whether these goals were worthwhile in the first place, or how different goals should be prioritised. To suggest that this could happen based on plausible research findings is naïve; basing a systematic evidence review on the premise of settling these priorities is a waste of time. Instead, *principle* recognises the role of ideology in policy-making: to choose a particular goal, or to prioritise one goal over another, is a normative choice.

Consider, for example, a systematic review of the evidence on the impact of corruption on economic growth, which was executed in 2011 and received funding from DFID (Ugur and Dasgupta, 2011). The question as to whether corruption is good, neutral or bad for economic growth is by no means invalid from a research point of view. The argument that corruption might be beneficial for growth by greasing the wheels for those engaging in high return activities, or by accumulating Schumpeterian rents that could spur innovation used to be very common, even dominant (Goldsmith, 2006). Whether and how the central research question of this systematic review can feed into the policy process is an entirely different matter. How would the policy process change, for example, if this systematic review concluded that corruption is beneficial for economic growth?[2] Would DFID start to design policies stimulating corruption? Since corruption is widely perceived as unfair, dirty or *unprincipled,* and constituting a reason to refrain from giving development aid, such a policy turn is highly unlikely.[3] Alternatively, DFID might attach a lower priority to policies countering corruption. Still the basis of such a decision would be a *principled*, not evidence-based, argument that the goal of stimulating economic development has priority over fighting the unfairness of corruption. Given that choosing and prioritising broad policy goals in this field is likely to be based on *principle*, systematic evidence reviewing is not useful in this context.[4] It is more likely that the evidence review would be used to justify policy

2. Two out of 28 studies considered in Ugur and Dasgupta (2011) draw this conclusion.
3. See, for example, the high profile 1.4 billion reasons campaign by the Global Poverty Project, which deals with corruption in some detail.
4. This is not to imply that the study by Ugur and Dasgupta is not useful; it lists evidence for and against numerous mechanisms connecting corruption and growth, which could be extremely useful when building a plausibility case for a specific policy to counteract these mechanisms. The point here is that this would be a different stage of the policy process: when the broad goal of fighting corruption has been established, and specific policies are sought in order to attain this goal.

decisions that have already been taken if conclusions are convenient (an example of policy-based evidence), and that the review would be ignored otherwise.

The example of food aid also illustrates how the emphasis on evidence-based policy can be unproductive when setting principled policy goals. There is a longstanding debate within the humanitarian and policy community as to whether food aid is a disincentive for local food production, and thereby hampers long-term development (see for example (Abdulai et al., 2005)). Regardless of the outcome of the academic debate on this adverse effect, whether this is a sound reason to stop food aid – potentially condemning individuals to starvation – is a question of principle.

We do not categorically say that setting broad policy goals using *principle* is a bad thing. A trade-off between corruption and growth, or relieving humanitarian need and long-term development, is of course driven by norms and beliefs. What is a bad thing to do is to pretend that these principled decisions are informed by evidence that can be systematically sought out to make them seem more grounded, thus silencing those who disagree with the principle. The *principle* debate needs to be based not on empirical evidence, but on what common ideologies can be used to design policies.

3.4 The visionary P: possibility

Something is still missing in the evidence picture we have presented up to this point. How can development policy innovate, move forward and discover new ways to achieve policy goals that are deemed worthy? Principle might help us to have a more honest discussion of the goals behind policy. Proof and plausibility emphasise finding out what has worked somewhere, and what could plausibly work in the future. But potential for innovation is limited: principle, proof and plausibility can only teach us something about policies that have been tried before. They can help us do fewer of the things that plausibly do not work, and more of the things that plausibly do work. But they cannot help us conceive of something new.

The notion of innovation seldom enters into the systematic review process because, in this context, the definition of evidence is commonly restricted to empirical information. This implies that no evidence can exist for a policy that is newly conceived. Taking this notion to the extreme, requiring all policy to be evidence-based by definition means no new policy can ever be tried. There is little room to be unapologetically big and bold, new or experimental. The fourth P allows for such room: *possibility*.

We posit that there is really no need to use any empirical information to develop big perspectives about change. Such perspectives often emerge from seminal works. These can be big think pieces that have the power to change our understanding of development, and they commonly do not achieve this with empirical evidence. Hence, they do not appear in evidence reviews. In fact, one of the major critiques directed at our evidence review was that we missed seminal works. Even though big ideas can suggest innovative policy, there is no obvious space to facilitate this in the current evidence landscape.

How can systematic evidence reviewing be more conducive to innovation? By allowing ideas to stand as evidence. A first step is not to automatically exclude non-empirical work. In the initial phase of selecting a pool of literature, systematic reviewers could include a selection of theoretical work or think pieces, selected on the basis of three criteria. First, the extent to which the work is recent, which could be easily determined through the publication date. Second, the degree to which the theory or idea is already covered by empirical work, which is presumably also found through the initial literature search. If this is the case, plausibility rather than possibility could take over. Third, the degree to which the work is well-respected, which can be roughly estimated by the ranking of the outlet it is published in, and/or the degree to which it is cited by others. In the second stage of the review, when the quality of empirical evidence is assessed, systematic reviewers could similarly assess theoretical work or think-pieces. They could evaluate whether the theory or idea presented is truly new, rather than a reformulation of already existing, empirically-researched ideas. Furthermore, they could assess the quality of the theory or idea itself. A sound theory needs to be internally consistent (meaning it confirms to basic rules of logic and does not contradict itself), does not make unrealistic assumptions that drive the theory (meaning if these assumptions are changed, the predictions from the theory change meaningfully) and makes a believable case that it explains some real-world phenomenon.

Innovation can only happen if we accept that trying out new things comes with associated risks. Currently, development agencies are quite averse to *programmatic risk*, the risk that a programme fails to achieve its objectives, or even cause harm. An OECD study of development programming in states affected by conflict finds, for example:

> In many cases development agencies have avoided high risk programming choices required to support statebuilding, peacebuilding and other forms of transformational change, and have instead opted for safer programmes concerned with direct service delivery. [...] These tendencies limit agencies' ability to address the challenges of statebuilding and peacebuilding. (OECD, 2014: 25)

Innovation can have a high return, but it is also risky. With Easterly, we argue for possibility within the evidence debate: for incremental change, for experimenting with innovative policies, and for recognition that when trying ten such policies, nine might be failures but this may be the price we have to pay to discover the tenth, high-return policy (Easterly, 2002).

Possibility as an approach changes the point of a systematic evidence review entirely: rather than seeking out empirical evidence, it moves the notion of evidence review towards a review of ideas and methods of investigation. What can then be presented is more akin to a narrative synthesis through an account of discourse and reflection, rather than through the pursuit of statistical and systematic analysis of existing information.

Possibility, our fourth perspective on evidence, is necessary for the recognition of emerging ideas and ways of seeing the world. To have exploration of new

possibilities through fresh perspectives crowded out by an evidence debate with a narrow technical focus means losing the motivation to search for evidence in the first place: to join a quest for a better understanding of the world, the people in it and what they collectively might be able to achieve.

4 Conclusion

We started off, somewhat grudgingly, by fulfilling a requirement to deliver a range of systematic evidence reviews to our donor. Soon we got hooked on thinking about what was right or wrong about them – how difficult it was to be systematic, how the focus on specific types of information might decrease the space for innovation, ignore seminal ideas and privilege out-of-context but gold standard empirical information over context-specific information of lower quality or theory.

Of even greater interest was noticing how some of these experiences linked to the much bigger debate about how to translate the findings or omissions of systematic reviews to usable information for development policy. No matter how much systematic reviewing on specific research one does, it does not automatically lead to finding out what the evidence landscape looks like. Even less obvious was how systematic evidence reviews might helpfully inform policy – after all the process of policy-making is in itself not that clear, having been described by Clay and Schaeffer as 'a chaos of purposes and accidents', or as 'complex' (Ramalingam et al., 2008), 'multifactoral' and 'non-linear' (ODI, 2009: 1).

There is much talk of a gold standard of evidence, but what to do with such precious metal remains rather unexplored. Thus, we propose abandoning the notion that the purpose of systematic evidence reviews is to determine quality of evidence. Rather, we put an ornate frame around different ways of seeing usability of evidence, which we call the 'Four Ps'. Each of these Ps represents a perspective on evidence that is useful in different stages or for different purposes of the policy process; none of these is usefully left out of systematic evidence reviews. *Principle* may be invoked when broad policy goals are set, a process that more often than not involves normative choices. A transparent discussion about which principles have a broad base of support is more productive than searching for fig leaves of evidence to justify predetermined policy in the eyes of those not subscribing to the principle underlying it. *Plausibility* comes in when considering alternative policies to attain broad policy goals, and we suspect this would be the perspective used most often. Recognition that even gold standard information does not guarantee future policy success allows us to include information commonly *not* considered evidence, such as lower quality studies or theory, to play a role in the policy process. Use of evidence as *proof* is relatively limited, although one can strive for the standard of proof when doing an evaluation. Finally, *possibility* brings innovation and experimentation back in, and in the process makes the job of conducting a systematic evidence review much bigger.

When systematically reviewing evidence, thinking of proof, plausibility, principle and possibility – rather than wanting to examine the quality of evidence – requires constant questioning of one's own perspectives on evidence. In translating

evidence to policy-relevant information, it means searching for the right perspective, redefining success and failure, and adjusting Theories of Change based on newly understood plausibility, which could possibly change perspectives entirely. This may be frustrating, it may be repetitive, it may not be streamlined and it is unlikely to be efficient. It might feel more like aluminium than gold; it will probably not feel as clean as the undertaking of a systematic evidence review suggests. Yet it is a way to constructively manage the encounters between solid social science and unpredictable and messy humans – whether these are researchers, policy-makers and practitioners, or those at the receiving end of international development policies. A less exclusionary approach to assessing evidence is necessary to at least enable international development policies to be made in the most informed way possible.

first submitted October 2014
final revision accepted June 2015

References

Abdulai, A., Barrett, C. B. and Hoddinott, J. (2005) 'Does Food Aid *Really* Have Disincentive Effects? New evidence from sub-Saharan Africa', *World Development* 33: 1689–704.

Adelman, S., Gilligan, D. O. and Lehrer, K. (2008) *How Effective are Food for Education Programs?*. Washington, DC: International Food Policy Research Institute.

Andreas, P. and Greenhill, K. M. (2010) *Sex, Drugs and Body Counts*. Ithaca, NY: Cornell University Press.

Blattman, C. and Miquel, E. (2010) 'Civil War', *Journal of Economic Literature* 48 (1): 3–57.

Carayannis, T.; Bojicici-Dzelilovic, V.; Olin, N.; Rigterink, A. and Schomerus, M. (2014) *Practice Without Evidence: Interrogating conflict resolution approaches and assumptions*. JSRP Paper 11. London: Justice and Security Research Programme, London School of Economics and Political Science.

Cartwright, N. D. (2007) 'Are RCTs the Gold Standard?', *Biosocieties* 2: 11–20.

Cartwright, N. D. and Bradburn, N. (2011) 'A Theory of Measurement', in R. M. Li (ed.), *The Importance of Common Metrics for Advancing Social Science Theory and Research: Proceedings of the National Research Council Committee on Common Metrics*. Washington, DC: National Academies Press.

Cartwright, N. D. and Munro, E. (2010) 'The Limitations of Randomized Controlled Trials in Predicting Effectiveness', *Journal of Evaluation in Clinical Practice* 16: 260–6.

Castillo, N. M. and Wagner, D. A. (2013) 'Gold Standard? The Use of Randomized Controlled Trials for International Educational Policy', *Comparative Education Review* 58(1): 116–73.

Collins, J. A. (2004) 'Balancing the Strengths of Systematic and Narrative Reviews', *Human Reproduction Update* 11(2): 103–4.

Cuvelier, J., Vlassenroot, K. and Olin, N. (2013) *Resources, Conflict and Governance: A critical review of the evidence*. JSRP Paper 9. London: The Justice and Security Research Programme, London School of Economics and Political Science.

Department for International Development (2014) *Assessing the Strength of Evidence*. London: DFID.

Department for International Development (2013) *Systematic Reviews in International Development*. London: DFID.

Duvendack, M.; Hombrados, J. G.; Palmer-Jones, R. and Waddington, H. (2012) 'Assessing 'What Works' in International Development: Meta-analysis for sophisticated dummies', *Journal of Development Effectiveness* 4(3): 456–71.

Easterly, W. (2002) *The Elusive Quest for Growth. Economists' adventures and misadventures in the tropics*. Cambridge, MA: MIT Press.

Farrington, D. P.; Gottfredson, D. C.; Sherman, L. W. and Welsh, B. C. (2002) 'The Maryland Scientific Methods Scale', in L. W. Sherman, D. P. Farrington, B. Welsh and D. MacKenzie (eds), *Evidence-Based Crime Prevention*. London: Routledge.

Forsyth, T. and Schomerus, M. (2013) *Climate Change and Conflict: A systematic evidence review*. JSRP Paper 8. London: The Justice and Security Research Programme, London School of Economics and Political Science.

Gasteen, M. (2010) *Systematic Reviews and Evidence-Informed Policy: Overview*. London: DFID Research and Evidence Division.

Goldsmith, A. (2006) 'Slapping the Grasping Hand: Correlates of political corruption in emerging markets', *American Journal of Economics and Sociology* 58(4): 865–83.

Grant, A. J. and Taylor, I. (2007) 'Global Governance and Conflict Diamonds: The Kimberley Process and the quest for clean gems', *The Round Table: The Commonwealth Journal of International Affairs* 93: 385–401.

Greenhalgh, T.; Robert, G.; Macfarlane, F.; Bate, P.; Kyriakidou, O. and Peacock, R. (2005) 'Storylines of Research in Diffusion of Innovation: A meta-narrative approach to systematic review', *Social Science and Medicine* 61(2): 417–30.

Greening, J. (2013) Development in Transition (Speech by the Development Secretary Justine Greening setting out her priorities for UK aid in the coming years, hosted by ONE Campaign UK).

Hannes, K. and Macaitis, K. (2012) 'A Move to More Systematic and Transparent Approaches in Qualitative Evidence Synthesis: Update on a review of published papers', *Qualitative Research* 12(4): 402–42.

House of Commons (2006) *Scientific Advice, Risk and Evidence Based Policy Making: Seventh Report of Session 2005–06*. London: Science and Technology Committee.

Luckham, R. and Kirk, T. (2012) *Security in Hybrid Political Contexts: An end-user approach*. JRSP Paper 2. London: The Justice and Security Research Programme, London School of Economics and Political Science.

Macdonald, A. (2013) *Local Understandings and Experiences of Transitional Justice: A review of the evidence*. JRSP Paper 6. London: The Justice and Security Research Programme, London School of Economics and Political Science.

Mallett, R.; Hagen-Zanker, J.; Slater, R. and Duvendack, M. (2012) 'The Benefits and Challenges of Using Systematic Reviews in International Development Research', *Journal of Development Effectiveness* 4(3): 445–55.

Mansuri, G. and Rao, V. (2003) *Localizing Development: Does participation work?*. Washington, DC: World Bank.

Mekasha, T. J. and Tarp, F. (2013) 'Aid and Growth: What meta-analysis reveals', *The Journal of Development Studies* 49(4): 564–83.

Miguel, E. and Kremer, M. (2004) 'Worms: Idenfitying impacts on education and health in the presence of treatment externalities', *Econometrica* 72(1): 159–217.

ODI (2009) *Helping Researchers Become Policy Entrepreneurs: How to develop engagement strategies for evidence-based policy-making*. London: Overseas Development Institute.

ODI Secure Livelihoods Research Consortium (2012) *Making Systematic Reviews Work for International Development Research*. London: Overseas Development Institute.

OECD (2014) *Development Assistance and Approaches to Risk in Fragile and Conflict-Affected States*. Paris: Organisation for Economic Co-operation and Development.

Paul, G., Kremer, M. and Moulin, S. (2009) 'Many Children Left Behind? Textbooks and test scores in Kenya', *American Economic Journal: Applied Economics* 1(1): 112–35.

Ramalingam, B.; Jones, H.; Reba, T. and Young, J. (2008) *Exploring the Science of Complexity: Ideas and implications for development and humanitarian efforts*. London: Overseas Development Institute.

Rodgers, M.; Sowden, A.; Petticrew, M.; Arai, L.; Roberts, H.; Britten, N. and Popay, J. (2009) 'Testing Methodological Guidance on the Conduct of Narrative Synthesis in Systematic Reviews: Effectiveness of interventions to promote smoke alarm ownership and function', *Evaluation* 15(1): 49–73.

Seckinelgin, H. and Klot, J. F. (2014) 'From Global Policy to Local Knowledge: What is the link between women's formal political participation and gender equality in conflict-affected contexts?', *Global Policy* 5(1): 36–46.

Snilstveit, B. (2012) 'Systematic Reviews: From 'bare bones' reviews to policy relevance', *Journal of Development Effectiveness* 4(3): 388–408.

Snilstveit, B., Oliver, S. and Vojtkova, M. (2012) 'Narrative Approaches to Systematic Review and Synthesis of Evidence for International Development Policy and Practice', *Journal of Development Effectiveness* 4(3): 409–29.

Stein, D. and Valters, C. (2012) *Understanding Theory of Change in International Development*. JSRP Paper 1. London: The Justice and Security Research Programme, London School of Economics and Political Science.

Stewart, R., van Rooyen, C. and de Wet, T. (2012) 'Purity or Pragmatism? Reflecting on the use of systematic review methodology in development', *Journal of Development Effectiveness* 4(3): 430–44.

Ugur, M. and Dasgupta, N. (2011) *Evidence on the Economic Growth Impacts of Corruption in Low-Income Countries and Beyond*. London: University of Greenwich.

Waddington, H.; White, H., Snilstveit, B. et al. (2012) 'How To Do a Good Systematic Review of Effects In International Development: A tool kit', *Journal of Development Effectiveness* 4(3): 359–87.

Wantchekon, L. (2003) 'Clientelism and Voting Behavior: Evidence from a field experiment in Benin', *World Politics* 55(3): 399–422.

White, H. and Waddington, H. (2012) 'Why Do We Care About Evidence Synthesis? An introduction to the special issue on systematic reviews', *Journal of Development Effectiveness* 4(3): 351–8.