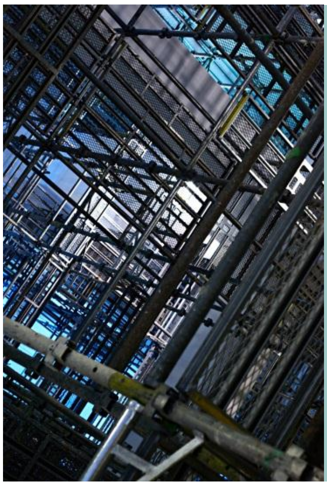
September 2013

Working paper



How to do a rigorous, evidence-focused literature review in international development

A Guidance Note

Jessica Hagen-Zanker and Richard Mallett

Abstract

Building on previous reflections on the utility of systematic reviews in international development research, this paper describes an approach to carrying out a literature review that adheres to some of the core principles of 'full' systematic reviews, but that also contains space within the process for innovation and reflexivity. We discuss all stages of the review process, but pay particular attention to the retrieval phase, which, we argue, should consist of three interrelated tracks – important for navigating difficult 'information architecture'. We end by clarifying what it is in particular that sets this approach apart from fuller systematic reviews, as well as with some broader thoughts on the nature of 'the literature review' within international development and the social sciences more generally. The paper should thus be seen as sitting somewhere between a practical toolkit for those wishing to undertake a rigorous, evidence-focused review and a series of reflections on the role, purpose and application of literature reviews in policy research.

Acknowledgements

The authors would like to thank the following for feedback on an earlier draft of this paper: Will Evans (DFID), Hugh Waddington (3ie), Maren Duvendack (ODI), Nat Mason (ODI), Jonathan Glennie (ODI) and Rachel Slater (ODI). Thanks also to Roo Griffiths for copy editing. The cover image is attributed to Jan Moren (license: CC BY-NC).

This paper is published jointly by the Overseas Development Institute and the Secure Livelihoods Research Consortium (SLRC). SLRC aims to generate a stronger evidence base on how people make a living, educate their children, deal with illness and access other basic services in conflict-affected situations (CAS). Providing better access to basic services, social protection and support to livelihoods matters for the human welfare of people affected by conflict, the achievement of development targets such as the Millennium Development Goals (MDGs) and international efforts at peace- and state-building.

At the centre of SLRC's research are three core themes, developed over the course of an intensive one-year inception phase:

- State legitimacy: experiences, perceptions and expectations of the state and local governance in conflict-affected situations
- State capacity: building effective states that deliver services and social protection in conflict-affected situations
- Livelihood trajectories and economic activity under conflict

The Overseas Development Institute (ODI) is the lead organisation. SLRC partners include the Centre for Poverty Analysis (CEPA) in Sri Lanka, Feinstein International Center (FIC, Tufts University), the Afghanistan Research and Evaluation Unit (AREU), the Sustainable Development Policy Institute (SDPI) in Pakistan, Disaster Studies of Wageningen University (WUR) in the Netherlands, the Nepal Centre for Contemporary Research (NCCR), and the Food and Agriculture Organization (FAO).

Table of contents

Figures

Abbreviations

AusAID	Australian Agency for International Development
CGD	Center for Global Development
DDR	Disarmament, Demobilisation and Reintegration
DFID	UK Department for International Development
GRADE	Grades of recommendation assessment, development and evaluation
IDS	Institute of Development Studies
MSMS	Maryland scientific measurement scale
NGO	Non-governmental organisation
PICO	Population, intervention, comparator and outcomes
ODI	Overseas Development Institute
RCTs	Randomised control trials
REAs	Rapid evidence assessments
SLRC	Secure Livelihoods Research Consortium

1 Introduction

Systematic reviews – a 'rigorous method to map the evidence base in an [as] unbiased way as possible, and to assess the quality of the evidence and synthesise it' (DFID, 2013a) – are considered by some to offer 'the most reliable and comprehensive statement about what works' (Petrosino et al., in van der Knaap et al., 2008: 49). Used widely and for many years in medical research and the natural sciences, systematic reviews have come to be seen as a key tool for evidence-informed policymaking within the arena of international development. As such, a number of bilateral donors – most notably the UK's Department for International Development (DFID) and the Australian Agency for International Development (AusAID) – have funded a series of systematic reviews over the past few years, with the express aim of finding out 'what works' in generating development outcomes.

However, a number of recent contributions to the development literature have called into question the framing of systematic reviews as a neutral, objective and comprehensive approach to evidence retrieval, grading and synthesis (Mallett et al., 2012; Walker et al., 2013; see also Newton et al., 2012 and O'Mara-Eves et al., 2013 for recent critical reflections from the fields of clinical psychology and public health, respectively). At the core of these critiques is the concern that systematic reviews, if carried out in a rigid and non-reflexive manner, may generate partial and misleading 'statements about what works' that are nevertheless seen to be authoritative and trustworthy.

As a direct response to this concern, this paper describes a way of carrying out a less rigid and more reflexive form of evidence-focused literature review. It is informed by our experiences of carrying out a series of systematic reviews in recent years, some of which have rigidly followed a set protocol, whereas others have taken a more flexible approach. The process outlined here is designed to produce a review strategy that adheres to the core principles of systematic reviews – rigour, transparency and a commitment to taking questions of evidence seriously – while allowing for a more flexible and user-friendly handling of retrieval and analysis methods. Such an approach is increasingly being used by researchers at the Overseas Development Institute (ODI), although not yet in any kind of standardised way (see Hagen-Zanker and Leon, 2013; Walker et al., 2013).

The approach described here walks the reader through each stage of a literature review process in a clear and useful way – a process that has been designed specifically to foreground the importance of taking empirical evidence seriously, to minimise retrieval bias and to ensure relevance and utility of the final product. We place particular emphasis on the need to get the retrieval phase right. As such, we propose a mechanism consisting of three interrelated tracks: academic literature search (Track I); snowballing (Track II); and grey literature capture (Track III). Using this mechanism and following the accompanying instructions provided below will help produce a focused review that captures material from a broad range of sources and locations – something considered particularly important in producing as comprehensive a review as possible (DFID, 2013b).

The paper follows a simple structure. In Section 2, we provide a brief discussion of the advantages and limitations of both orthodox literature reviews and systematic reviews. We then describe our approach to carrying out a rigorous, evidence-focused literature review (Section 3), outlining the three tracks involved. Finally, Section 4 concludes by highlighting what it is that sets this particular approach apart from ‘full’ or ‘official’ systematic reviews.

2 Literature reviews in international development: from orthodox to systematic

This section situates the discussion on literature review methods and states the problem being addressed through this paper. It is split into two short subsections: the first characterises, in a stylistic way, the typical nature of orthodox literature reviews in international development; the second explores the use of systematic reviews in international development in greater depth. We draw throughout this section on previous discussions presented in Hagen-Zanker et al. (2012) and Mallett et al. (2012).

2.1 Shortcomings of orthodox literature reviews

Orthodox literature reviews involve a review of the literature on a given subject in order to generate an answer to a specific research question. Based on our experience, such reviews typically start with the studies authors already know in a particular area – which tend to be highly cited, seminal pieces of work. These are then complemented by resources identified – not necessarily systematically or transparently – through various channels, including academic databases, individual peer review journal archives, institutional websites and internet search engines (such as Google). While this makes for a relatively quick and easy process, it is characterised by a number of shortcomings.

The main one is that there is a strong bias in selecting literature; put crudely and stylistically, we all start with the authors and literature we know, we are more likely to use the literature we can access easily and we are more likely draw on those studies that have significant (and preferably positive) findings. This means orthodox literature reviews may draw conclusions from what is essentially a 'non-representative sample' (Petticrew and Roberts, 2006).

Another shortcoming lies in the actual review phase, where data extraction and analysis of retrieved material may be carried out in a variable and non-transparent way (Petticrew and Roberts, 2006). Generally speaking, orthodox literature reviews do not adopt a predefined approach to assessing or grading evidence, and tend to be far more concerned with the results of retrieved studies than with questions of research design, methods used in the generation of data and the quantitative and qualitative dimensions of datasets drawn on.

What this ultimately means is many orthodox literature reviews may not represent trustworthy products[1]. If one cannot be sure that a review comprehensively and accurately portrays and assesses – in an unbiased way – the literature on a particular subject, then it is difficult to accept the review's findings as robust. Of key importance here is the issue of transparency in the review and write-up process – something at the core of systematic reviews.

2.2 Systematic reviews in development studies

Owing in part to the shortcomings of orthodox literature reviews outlined above, systematic reviews have recently been introduced into the field of international development, having been used in the natural sciences for decades already.

Systematic reviews are essentially a rigorous and transparent form of literature review (see Waddington et al., 2012 for a toolkit on how to do a systematic review in development studies, and Mallett et. al., 2012 for a discussion on the steps generally involved in systematic reviews). Systematic reviews closely adhere to a set of predefined standards and steps in identifying, assessing and synthesising evidence in order to limit bias.

Adhering to core systematic review principles — rigour, transparency, replicability — can improve the quality and strength of orthodox literature reviews in a number of ways: first, by increasing the breadth in the literature included, while retaining focus; second, by focusing on empirical evidence, not preconceived knowledge; and third, by being transparent and replicable. When these principles are applied sensitively, systematic reviews have a clear advantage over orthodox literature reviews: the quality of reviews is improved through greater transparency; a wider range of studies are identified and screened; implicit researcher bias is theoretically reduced; and reviewers are encouraged to engage more critically with the quality of the evidence.

However, systematic reviews are difficult to apply in practice, and raise a number of practical challenges and methodological concerns (see Box below).

Practical challenges and methodological concerns associated with systematic reviews

Practical challenges

Systematic reviews require access to a wide range of databases and peer-reviewed journals, which can be problematic and very expensive for non-academic researchers and those based in Southern research organisations.

Searching institutional websites, for example those of international organisations, is essential to ensure breadth of systematic reviews, as relevant research is often located outside formal peer-reviewed channels. However, searching institutional websites undermines the objectivity of the search and retrieval process and introduces bias to the review process.

In order to achieve objectivity, inclusion and exclusion criteria are used to screen potentially relevant studies. However, there is inevitable subjectivity in the screening process, particularly when high numbers of researchers are involved, as each member of the research team interprets inclusion criteria slightly differently,

[1] We are using the term 'trustworthy' here in line with its methodological application in the social sciences (see DeMarrais and Lapan, 2004).

introducing another source of bias.

Methodological concerns

There is an inherent contradiction between the information required to conduct a systematic review and the way peer-reviewed journal articles are written in development studies. Empirical impact studies in development studies are not written in a uniform fashion, unlike in the natural and medical sciences or even compared with economics. This is problematic from a practical perspective, but also means the attributes that get research published in a peer-reviewed development journal are very different to those required for inclusion in a systematic review.

It may be much harder to assess evidence ‘objectively’ in development studies compared with in other fields, such as those in which systematic reviews were pioneered.

Systematic reviews often miss context and process, understandings of which are central to good international development (as well as broader social science) research.

Sources: Hagen-Zanker et al (2012); Mallett et al (2012).

In order to overcome some of these practical challenges and methodological concerns, we propose an alternative approach to carrying out a systematic review, one that adheres to the core principles of ‘full’ systematic reviews but allows for greater flexibility and reflexivity in the process.

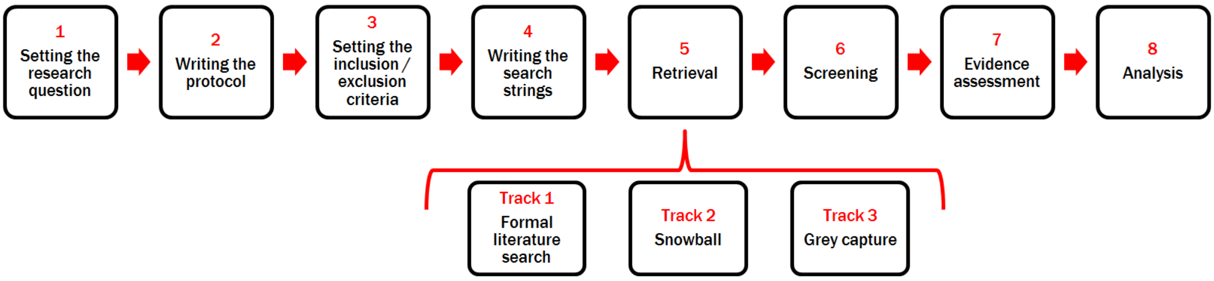
The review process described in the next section is similar to that used in Rapid Evidence Assessments (REAs), which are defined as a tool for finding out what is known about a particular issue drawing on systematic review methods (Civil Service, 2012). REAs place great emphasis on a rigorous search process, but tailor the approach for those facing time and resource constraints. However, while the REA process still fairly closely resembles a systematic review – demanding, for example, that at least a simple quality assessment of the evidence is carried out – the approach described in this paper incorporates greater room for flexibility and iterative problem solving, the aim of which is to help the review team ensure the final product of their efforts is both relevant and usable.

3 What the process looks like and how it works

Rather than following a rigid systematic review methodology, our shared experience suggests a more useful approach for development researchers might involve a mixture of compliance and flexibility: compliance with the broad systematic review principles (rigour, transparency, replicability) and flexibility to tailor the process towards improving the quality of the overall findings, particularly if time and budgets are constrained. Having incorporated this mixture into the design of some of our own approaches to systematic literature review, we describe below the outcome of our experimentation.

This section is intended to walk researchers through the review process in a straightforward and helpful way, providing advice and pointers where appropriate and flagging issues to consider when necessary. Subsequently, we structure the section in a kind of chronological order, following the stages outlined in Figure 1 below and paying particular attention to the multi-tracked retrieval mechanism in Stage 5.

Figure 1: Stages in a rigorous, evidence focused literature review


1 Setting the research question
2 Writing the protocol
3 Setting the inclusion / exclusion criteria
4 Writing the search strings
5 Retrieval
6 Screening
7 Evidence assessment
8 Analysis
Track 1 Formal literature search
Track 2 Snowball
Track 3 Grey capture

Source: Authors' adaptation of existing approaches

Stage 1: Setting the research question

Setting a feasible research question is the starting point for any review.[2]

When deciding, it is important to think carefully about both the focus of the question and the specific phrasing and language to be used. Keep the following guidance in mind when setting the question:

- Stick to one question (rather than trying to squeeze too much into one question).
- Make the question as specific as possible.
- It helps to phrase question in terms of 'What is the evidence on ... ?'
- It helps to think about the question in terms of *population*, *intervention*, *comparator* and *outcome(s)* (in short, PICO), particularly if researchers are interested in questions around impact. Breaking down the research question in this way is useful for clarifying the focus:
  - *Population* – Who are we looking at? One particular subgroup within a given population? Do we restrict the search to specific countries (e.g. fragile and conflict-affected situations)?
  - *Intervention* – Are we looking at the impact of a particular kind of programme/event/change that took place (e.g. a disarmament, demobilisation and reintegration (DDR) programme; the involvement of women in a peace-building process)?
  - *Comparator* – What comparison group (if any) are we using to compare the outcomes of the intervention? Comparison groups can be constructed on the basis of different social groups, different geographical contexts, different time periods, and so on (e.g. what are the impacts of a cash transfer: on male vs. female household members; in conflict-affected vs. stable settings; from 1990-2000 vs. 2001-2010?)
  - *Outcome(s)* – Specify which outcome or set of outcomes we are interested in as a result of the intervention (e.g. impact on poverty). Note that outcomes and impacts are not the same as outputs, which are more concerned with the internal success of a programme (e.g. x combatants were disarmed; x mosquito nets were distributed; x microfinance loans were allocated).
- Expect to revise the research question a number of times. For example, if it turns out that the research questions results in too few/too many search hits, the question may need to be broadened/reduced in scope.

It should also be noted that, although systematic reviews in international development have tended thus far to focus on questions around impact, this is not the only way of framing research questions for evidence-focused literature reviews. We might, for example, be more interested in examining the transmission mechanisms underpinning change, in asking about the process of programme implementation rather than simply whether positive impacts were produced.[3]

[2] Waddington et al. (2012) provide a good discussion on how to set an appropriate research question for a review.

[3] See Civil Service (2012) for detailed guidance on the different kinds of questions that can be asked.

Stage 2: Writing a protocol

Previous experience suggests it can be helpful to write out the search strategy in a protocol format before searching is started. The protocol should be as clear and straightforward as possible, and can be referred back to throughout the search process to ensure all members of the research team are taking the correct steps. Writing a protocol also helps ensure transparency and will enable similar steps to be taken in the future if replication is desired.

A protocol should include the following elements:

- A description and explanation of the research questions;
- An outline of the research methodology;
- A list and explanation of the inclusion and exclusion criteria;
- A list and explanation of the search strings;
- A list of the databases, journals and websites to be used;
- An explanation of how the data will be analysed and presented;
- A timeline.

In order to ensure the protocol is both thorough and rigorous, it can be helpful to have it peer-reviewed by someone familiar with systematic reviews or other forms of evidence-focused review.

Stage 3: Setting the inclusion/exclusion criteria

Establishing specific inclusion and exclusion criteria at the beginning of the process makes it easier to identify relevant material for review. It also helps improve both the transparency and the rigour of the review, by ensuring the screening is conducted in a consistent manner.

- Define which interventions are included (e.g. social protection instruments included in this search are cash transfers and public works programmes, but not health insurance).
- Specify the kind of study you want to include in the review: you need to be clear whether you are interested in studies that use only quantitative methods, only qualitative methods or both. Alternatively, you might be interested only in specific methods, such as randomised controlled trials (RCTs) or semi-structured interviews.
- Specify which outcome indicators will be considered (and which will not). For example, for nutrition, you might be interested only in weight-for-age and not height-for-age. Alternatively, you might decide to include all nutrition indicators.
- Decide on languages covered: this will depend on the scope of the study and the language skills of the research team, but it is important to agree which languages will be covered in the review. If the review covers multiple languages, a different protocol will need to be compiled for each language
- Decide on the time limit: is there a time cut-off for studies included? For example, if you set the limit at 1994, then studies published before this date must be excluded. Something to consider when deciding on a cut-off date is the population segment of the research question. For example, if you are focusing on countries officially defined as ‘fragile and conflict-affected’, then you may select your countries by using one of the many lists available (e.g. the Failed States Index, the World Bank Fragile Situations List). Bear in mind that any number of these countries may not have been classified as such x years ago. This means you may end up including countries that may well be

‘fragile and conflict-affected’ today but at the time of the study’s publication did not fall into this category. (The opposite scenario is also possible, where countries that were conflict-affected x years ago but are now not classified as such will likely be excluded.)

Stage 4: Writing the search strings

Getting your search strings right is extremely important, because your selections will determine what material you retrieve. As such, good testing of search strings should be factored into the process.

- Search strings need to include the relevant keywords related to intervention and outcome (and possibly population).
- If there are synonyms for the keywords or different kinds of interventions used, they need to be included in the search string (to ensure breadth in the search).
- The search string should use the most commonly used synonyms.
- If including the population in the search string, think carefully about how you want to define your groups of interest. Using the phrase ‘fragile and conflict-affected situation’ may not generate many hits, but be aware that using country names can also be problematic. For example, some lists may define Uganda as ‘fragile and conflict-affected’, but this applies only to a particular part of the country. Therefore, you may end up with many studies that fulfil all the inclusion criteria but tell you nothing about the conflict specificity of the case.
- It is important to carefully consider logical operators for Boolean searches: in some search engines, the logical operators AND, OR and NOT can be used to combine keywords in search strings. If two different logical operators are combined, the order of keywords needs to be thought through (and also the use of brackets). Different databases work slightly differently, so this needs to be tested every time.
- Search strings should be tested. If the search results in too many/too few hits (or do not include relevant studies), it is advisable to revise the search strings. In terms of which side to err on, it may be better to use search strings that generate too many results, because irrelevant studies will be excluded during the screening process anyway (using the inclusion and exclusion criteria).

In addition to the above, a recent contribution from Thompson et al. (2013) discusses how certain kinds of software can be used to assist with search term selection for systematic reviews.

Stage 5: Retrieval

Through our own experience, we have realised that getting hold of relevant material is not only a vital stage in the review process – after all, it is the material gathered here that will help us answer the research question – but also a stage that throws up many challenges to overcome.

The ways documents tend to be stored in the field of international development, or the places they tend to be located, are not conducive to orthodox search and retrieval methods centred on academic journals and databases. There is often a great deal of relevant material that cannot be found within the traditional peer review storage system – for example working papers from think-tanks such as ODI, the Institute of Development Studies (IDS), the Center for Global Development (CGD) and so on; World Bank publications; non-governmental organisation (NGO) materials. Indeed, our experience of conducting systematic reviews suggests the majority of included studies are sometimes found through these alternative channels (Hagen-Zanker et al., 2012). As such, the problematic nature of what we might call the 'information architecture' within the field requires a more innovative approach to gathering relevant material.

Other have also raised questions as to the ability of rigid systematic reviews to identify relevant research: O'Mara-Eves et al. (2013), for example, reflect on a systematic review of the impacts of community engagement in public health interventions, estimating that typical searching and screening methods would have missed more than a quarter of the studies identified. If nothing else, this speaks to the importance of getting the retrieval phase 'right' in the first place.

Balance is the key here. You want to be able to search a broad range of platforms in order to identify all relevant information, but at the same time you want to minimise duplication of hits. We propose a retrieval mechanism consisting of three separate yet interrelated tracks: academic literature search; snowballing; and grey literature capture.

Track I: Academic literature search

Of the three tracks discussed here, Track I most closely resembles the typical procedure of plugging predetermined search strings (Stage 4) into academic databases in order to identify potentially relevant material. Although not sufficient in itself, it remains a necessary and important part of the retrieval process.

For the academic literature search, consider the following questions and issues before selecting your databases and journals to search.

- Which academic databases do you want to search and do you have access? Which fields do they specialise in?
- Are there any specific journals you want to search? It may be sensible to draw up a list of journals relevant to your subject area before you start.
- In order to avoid duplication, you can check if databases already search relevant journals.
- Are there any specific institutional websites you want to search?
- Decide if you are going to be setting limits on the number of studies to be reviewed if the search results in an excessively high number of hits.

Where to do the searches and access databases and academic journals can be a difficult question for anyone outside a (Western) academic institute. Usually, this means cooperating with someone from an academic institution, using an open access library or, depending on the budget of the review, possibly purchasing temporary access to academic journals.

Track II: Snowballing

Snowballing is the second track of the retrieval mechanism, and does not require the use of predetermined search strings. This process involves actively seeking advice on relevant publications in a particular field or on a particular topic from key experts – which will then be reviewed – and subsequently looking at the reference lists of those publications. This track is likely to introduce subjectivity to the process and a researcher bias to the studies included. As discussed at length earlier on, both the selection of key experts, and in particular their recommendations, will not be 'objective'. However, it can still be helpful to pursue this track, for example to get hold of non-published studies. Furthermore, this track is extremely useful to get a sense of which literature has been important and influential in the field – which may not necessarily be the high-quality peer-reviewed journal articles!

Snowballing involves the following steps.

1. *Identifying experts*

   To identify key experts, sit down with your team and together agree on a list of key experts working in that specific field. Depending on the research question, you need to ensure geographical variation and institutional variation (including both policy and academic experts). As a rule of thumb, we suggest five key experts – although this number is unavoidably arbitrary.

   In our experience, there is a dearth of evidence by researchers and organisations from the Global South, especially in formal channels. The snowballing track is an opportunity to counterbalance this and to actively seek out evidence emanating from the Global South, for example by including experts from Southern organisations.

2. *Identifying publications as starting points*

   Contact experts and ask them to suggest five or ten key publications on that particular question; also look at their publications and websites. Be clear on what kind of publications you are asking them to suggest. They do not necessarily have to be the highest-quality publications they know of, but instead may be considered particularly influential or widely cited.

3. *Snowballing*

   Look at the reference lists of those publications and look for other relevant publications on the same research question (using the inclusion/exclusion criteria can help in deciding on the relevance of publications). Then download those references and, possibly, look at the reference lists of those references. This is known as the 'backward snowball'. You can also do the 'forward snowball', which involves searching for studies that reference the studies recommend by experts.

Track III: Grey literature capture

As mentioned above, relevant material is often located outside the orthodox peer review channels (that is, academic databases, journals). Failing to incorporate a way of retrieving this material into the search strategy means you are unlikely to capture all available research – in particular research referred to as grey literature, such as working papers, concept notes, donor reports, policy documents and briefings. By including a list of institutional websites in the search strategy – as suggested above – it is possible to capture some of this material. But experience also suggests that using internet search engines, such as Google, can unearth further relevant material. This can be particularly helpful for identifying studies and reports that have just been released and would not therefore be picked up through snowballing. GoogleScholar can be a particularly effective way of identifying new or grey literature, and also features a useful 'cited by' function, which enables you to see where relevant studies have been cited (this can be considered an extension of the snowballing method – the forward snowball as described above).

More broadly, by incorporating Track III into the retrieval mechanism, it is possible to move away from some of the rigidity associated with systematic reviews (which let you include only material sourced through specific channels and mechanisms). Although often considered to be of lower quality than the peer-reviewed literature, a focus on grey literature can really help increase the breath, relevance, topicality and ultimate utility of your review.

Stage 6: Screening

Assessing the relevance of the literature is called screening. It is helpful to download the publication details and abstracts (and eventually full texts) and put these details into a reference manager (e.g. Mendeley) or specialist software such as EPPI-Reviewer or a Word/Excel document. This makes the screening stage easier, particularly it is often different people who do the downloading and screening.

When running the searches, reviewers tend to go through two rounds of screening. The first round is done on the basis of all the studies that come up using the agreed-on search strings. All these studies are then assessed on the basis of their titles and abstracts, using the inclusion and exclusion criteria. Studies need to meet all inclusion criteria to be included as relevant. On the other hand, meeting one exclusion criterion is enough to exclude a study from the review. For those studies that meet the inclusion criteria, or where insufficient information is available in the abstract to assess relevance, the full text will be downloaded. The second round of screening is then based on the full text of the documents, using the same inclusion/exclusion criteria. While the first stage of screening is often done by a research assistant (who should err on the side of inclusion if in doubt),[4] the second and final stage of screening should be done by the thematic expert. The studies that remain after the second stage of screening are those that are included in the review (in addition to studies found in the other tracks; see below).

Keep in mind that, depending on the breath of the question and the field, it is not unlikely to have hundreds or even thousands of search results that need to be screened. It is extremely important to assess all search results and studies consistently, so if a search results in a high number of hits, search result #1 needs to be assessed in the same way as search result #1001, which can only be done if time is budgeted for realistically. It is important to consider the possibility of having many search results and to decide beforehand how to approach this worst-case scenario (see Shemilt et al., 2013 for a discussion of how text mining technologies can help reviewers deal with impractical screening workloads). Essentially, either sufficient time and funds need to be set aside to screen all search results, or the research question needs to be narrowed down and the search rerun.

Please note that, while full systematic review policies tend to require researchers to collect and store details on the excluded studies, a lighter-touch approach may not ask researchers to do this. Further, unlike in full systematic reviews or REAs, the approach outlined here does not require researchers to record number of hits per database or number of irrelevant studies. We have decided not to collect this information, in order to devote more time and report space to the findings of the relevant studies rather than documenting the steps involved in the process.

Stage 7: Evidence assessment

Once the search has been completed, and the retrieved studies have been assessed using the inclusion and exclusion criteria, the relevant studies need to be described and classified. This information will also help assess the breadth and depth of the evidence, that is, the strength of the body of research.

This process can be tailored to your study and will be different for different fields. Assessing the strength of a body of research consists of three different components: 1)

[4] If screening is done by a research assistant or a number of researchers, it helps pilot screening, e.g. to screen the same 20 studies and then compare results. If there are any discrepancies in how they were screened, these should be discussed and agreement should be reached on the interpretation of the inclusion/exclusion criteria and on which studies should be included.

classifying and describing studies; 2) assessing the research quality of every study; and 3) assessing the overall strength of the body of research on the basis of the other two.

More specifically, the three components include the following steps:

1. *Classification of studies found*
   - It is helpful to include a table that describes the relevant studies.
   - Such a table should include basic descriptive information, such as year published, time period the study refers to, geographical coverage, intervention, outcome indicator, research design, research method.
   - Studies can be discussed in more detail by looking at a subset of studies (e.g. those using a particular outcome indicator, those focusing on a particular geographical area etc.).

2. *Assessing research quality*
   - This is notoriously difficult and inherently subjective. Here, it is of great importance that you be very transparent about which criteria are used and the justification for those criteria (see Box below).
   - Criteria generally include assessment of appropriateness and rigour of research design and method, reliability, validity and openness and transparency of the study (see also Box below and DFID, 2013b).
   - As an absolute minimum, we recommend identifying the data source and research method used. This should also be part of classifying studies.

3. *Assessing the overall strength of body of research*
   - Again, this can be quite subjective, and you will need to be clear about how you assess/describe this.
   - Following DFID (2013b), the following attributes may be considered and discussed: quality of the studies included in the body of evidence (e.g. median quality; distribution of evidence), size of the body of evidence (how many studies you found); consistency of the findings (do they come to the same conclusion?); and context of the evidence (global, regional or context-specific).
   - You will need to decide how the data will be analysed and presented. Mostly, it will be in the form of a table, but some of the information could be presented graphically.
   - Assessing and presenting the research evidence found is the first step in synthesising the research evidence. The same procedure can be repeated for subgroups of the overall body of evidence.

In our experience, Steps 2 and 3 – assessing research quality and the overall strength of the body of research – are notoriously difficult and inherently subjective (see also Box below). Even following seemingly clear and objective guidelines, such as those outlined in DFID (2013b), is neither a straightforward nor a bias-free exercise. For example, any answers to the question of whether a particular research method is appropriate to answering the research question will involve a fairly large degree of subjectivity; even within a fairly narrow research field there will always be some disagreement on which research methods are most appropriate depending on, say, the researcher’s background. Further, ranking evidence exclusively on the basis of the methods used to generate it presumes that variations in research quality are primarily determined by method. However, one might reasonably argue that such variation may be caused less by differences *between* methods and more by differences *within* methods. Two RCTs looking at the same research question, for example, may actually look very different, in terms of design, implementation and interpretation of results. Why should we place both on the same level, classifying them as high quality, when there may in actual fact be striking differences in the quality of each study?

Similarly, and to put it somewhat crudely, rigorous experimental methods, such as RCTs, are usually designed to achieve high internal validity. This is of course important if we are interested in establishing causality and attribution effects *in a particular* experimental study. But it is unclear what the (internally valid) findings produced by such studies tell us about what might be effective in another place at another time. (It is likely to be extremely limited.) Indeed, in a valuable recent exercise by Lant Pritchett and Justin Sandefur (2013), the authors attempt to measure bias from a) inferring causation from less rigorous, non-experimental estimates, and b) transferring more rigorous, internally valid findings from a single experimental study (an RCT) to different contexts – and then compare the size of the errors. They find that the error implied by generalising RCT results to other contexts is 'far, far greater' than the error implied by reading causality from studies that cannot actually claim to be causal. What does this mean? First, it speaks (loudly) to the dangers of taking highly rigorous experimental findings from one particular place and at one particular time, and claiming that they will apply in other situations. Given that systematic reviews encourage the ranking of evidence on the basis of methodology – and that they tend to privilege randomised, experimental methods achieving high internal validity – this is somewhat at odds with the *raison d'être* of systematic reviews: to inform decision makers about 'what works' across contexts. Second, it raises important questions about the way in which we grade evidence, and the multiple potential ways there are of doing this. Of course, impact findings generated through experimental methods are an important part of understanding what interventions work, but – on the basis of what we have just discussed – whether they are objectively 'better' or 'more helpful' than findings generated by non-experimental and less quantitative methods is highly arguable (more on this in the Box below).

On a more practical note, being able to competently carry out any of the three classification and assessment steps outlined above depends on there being sufficient methodological information in the documents under review. Experience suggests this is not always the case, even in studies that have met all the initial inclusion criteria. This might be for a number of reasons. For example, authors publishing in peer-reviewed journals face strict limits on word count, meaning discussions of research content and findings might get prioritised over lengthy accounts of methodological approaches and practices (Hagen-Zanker et al., 2012). Whatever the cause, this is an issue that poses particular problems for appraisal; reviewers will have to decide whether to keep studies that discuss methodology in passing but simply do not provide enough detail for proper classification and assessment.

Following this discussion, our approach recommends stopping short of carrying out a full assessment of research quality, advocating instead for a stronger focus on conducting a comprehensive and useful classification of studies under review (see also Stage 8), which should – in theory – be more straightforward, more feasible and less time-consuming. We are not alone in advising this kind of approach. Recently, for example, Pritchett and Sandefur (2013: 34) concluded the following:

> *Avoid strict rankings of evidence. These can be highly misleading. At a minimum, evidence rankings must acknowledge a steep trade-off between internal and external validity. We are wary of the trend towards meta-analyses or 'systematic reviews' in development, as currently sponsored by organizations like DFID and 3ie.*

Notes on assessing the quality of evidence

Much has been written about the quality of evidence and how to assess it, within both international development and the broader public policy arena. Indeed, many would agree this issue is currently one of the 'hot potatoes' of development. As stated in the introduction to this paper, our intention here is not to advise substantively on the thorny question of how to assess the quality of evidence; indeed, this would require a separate, far longer paper (and it is not entirely certain that we would be able to do that discussion justice, even with a longer word count!) However, it is worth flagging up a couple of key issues to those considering undertaking a rigorous, evidence-focused literature review of their own.

First, scales for assessing the quality of evidence do exist, and they are relatively easy to find. The Grades of Recommendation Assessment, Development and Evaluation (GRADE) approach and the Maryland Scientific Measurement Scale (MSMS) represent just two scales (see also DFID, 2013b). However, reviewers need to think carefully about which scale(s) they opt for, and why – if, indeed, they do at all. Such scales are not neutral or apolitical; they are built on particular assumptions about the ordering of evidence, and, as such, it is important to question their foundations. Many, if not most, scales classify large RCTs as the 'best' kind of evidence, but the debate on the utility and appropriateness of these methods continues to rage with a surprising degree of ferocity. This debate cannot be treated as unrelated to questions around how to conduct a systematic review, carrying as it does significant implications for the way we construct hierarchies of evidence.

Second, generally speaking, social scientists seem more comfortable when it comes to assessing quantitative evidence. This is not to say the scales used for this purpose are not problem-free (see above point), but there is something about the overtly numerical dimension of quantitative evidence that seems to lend itself more easily to being graded. On the other hand, there appears to be a greater degree of apprehension when it comes to assessing qualitative evidence. It is possible to use a study's 'trustworthiness' as a proxy for rigour, but given that the main concern here is with the level detail provided in the text, is this perhaps setting the bar a little low? (As in: so long as a study contains an extensive discussion of methodology, data and theories of change, are we content to 'grade it well'?)

Readers may find it frustrating that we are unable to provide clear-cut answers to the kinds of questions raised here. However, this is an ongoing (and often divisive) debate, and all we can (or should) advise is that researchers remain abreast of developments, as well as approach quality assessment frameworks and scales with a critical and informed eye, keeping in mind that there is no objective, bias-free way of assessing evidence.

Stage 8: Analysis

In the last stage of the process, findings are described and summarised and synthesised with the aim of answering the overall research question. Unlike in full systematic reviews, the focus is solely on included studies, meaning excluded studies are not discussed as part of the synthesis and analysis. There are a number of different synthesis methods to synthesise the research findings and, to a large degree, the choice of synthesis method depends on the size of the relevant research body and the nature of the studies (e.g. depending on questions such as are they quantitative or qualitative? Are similar methodologies followed?) Synthesis methods include meta-analysis (using statistical methods to summarise and compare the findings) and narrative synthesis (describing and comparing the findings in great detail) (see Bergh, 2012 for a detailed description of different synthesis methods).

Carrying out a meta-analysis of included material is often considered a key part of systematic reviews. However, this can be both an extremely resource-intensive exercise and academically challenging, particularly when the sample of included studies demonstrates a high degree of methodological variation. In many cases, a statistics-based form of meta-analysis may be neither possible nor appropriate. Researchers may then instead consider narrative synthesis, which refers to 'an approach to the systematic review and synthesis of findings from multiple studies that relies primarily on the use of words and text to summarise and explain the findings of the synthesis' (Popay et al., 2006: 5). Narrative synthesis can be applied to a wide range of research questions, not only those concerned with programme effectiveness or impact, and may therefore be appropriate to exploring questions of how and why certain things work and others do not. A focus on impact, which seems to characterise many of the recently funded systematic reviews in international development, can sometimes lead to a preoccupation with finding out whether a particular outcome is being achieved by a particular programme, without exploring the (arguably more interesting and important) questions of how and why. Of particular relevance here are questions of external validity – that is, the extent to which the findings of a study can be legitimately transferred from one context to another. As Pritchett and Sandefur (2013) point out, the design or practice of many systematic reviews is often not appropriate to deal with this complex problem:

> *When non-experimental estimates vary across contexts, any claim for external validity of an experimental result must make the assumption that (a) treatment effects are constant across contexts, while (b) selection processes vary across contexts. This assumption is rarely stated or defended in systematic reviews of the evidence.*

They continue:

> *We conclude with recommendations for research and policy, including the need to evaluate programs in context, and avoid simple analogies to clinical medicine in which 'systematic reviews' attempt to identify best-practices by putting most (or all) weight on the most 'rigorous' evidence with no allowance for context.*

Thus, if we are interested in producing a review output that is helpful and relevant for practitioners, then we may need to explore different ways of integrating discussions of causal mechanisms and theories of change into our reviews, which might help explain some of the findings we see emerge. This, again, is an argument for adopting a review approach that complies with certain core systematic review principles, but also leaves enough space for iterative and flexible problem solving. The vignette presented in the Box below is one example of how this might be done in practice.

An example of how to use an iterative problem-solving approach to increase the utility of a systematic review

In 2011/12, Secure Livelihoods Research Consortium (SLRC) researchers carried out a systematic review of the impacts of seeds-and-tools interventions in fragile and conflict-affected situations. After the retrieval and screening phases, the review team was left with nine studies of relevance and merit. Although small in size, this sample was characterised by considerable heterogeneity in terms of contextual focus, specific programming type, methodology and assessed outcomes, making the synthesis and analysis of impacts both problematic and arguably undesirable. Rather than produce an output that focused purely on the question of whether seeds-and-tools programmes produced positive or negative impacts on a set of outcome indicators, the review team revisited the studies to consider whether there was anything that could be said beyond statements on impact. A broader, less constrained re-examination of the studies revealed often extensive discussions by the various authors of the process of programme implementation – that is, specifics of the design, circumstantial factors, assessments of how well programmes achieved internal outcomes and so on. While the review team was still unable to draw conclusive lessons on programme effectiveness, the revised focus – which was possible only because of an iterative and flexible review design – enabled the team to *say something* of value about how the nature of programme design and implementation might have important mediating effects on impact.

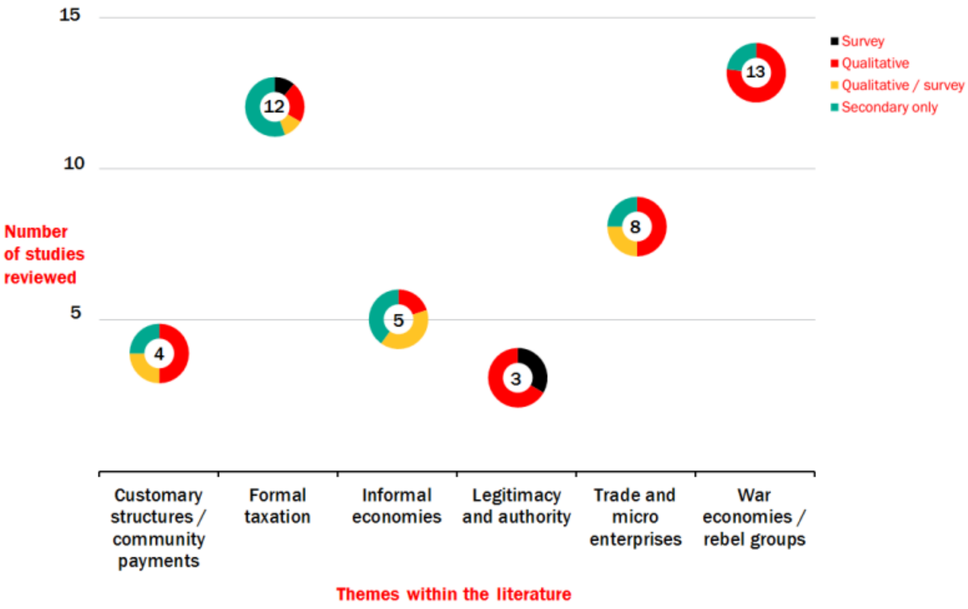
Other considerations to think about when working through the analysis stage of a review include:

- **Striking the right balance.** It can sometimes be difficult to write up analysis in a way that balances broad synthesis and summary with empirical and contextual detail. This is not to say the two are mutually exclusive, but rather doing both well in the same document can be quite challenging. Researchers need to think about the particular needs of their target audiences: will it be important to explore the details of each and every study reviewed, or will it be sufficient to paint a broad picture that, while lighter on the detail, may be much more accessible to potential readers? A recent exchange with a donor colleague confirmed the need to think carefully about the weighting of a review's discussion: 'I have noticed with some of my own work that the quest to review as much of the literature as possible (and then synthesise it) can make the written output a bit stodgy. Sometimes people come back and say, "It's great to have all the detail, but it's hard to see the wood for the trees." Obviously that's a challenge that the good analyst just has to overcome […] but I think it's why standard literature reviews retain their popularity.'

- **Commenting on the nature of the evidence.** One of the central features of a rigorous, evidence-focused literature review is that it goes beyond a discussion of simply what the evidence says – it should also have something of value to say about the nature of the evidence base itself. It is the job of the review team to give the reader a sense of what the evidence base looks like in terms of its quantity and quality, so it is possible to know what the review's findings are based on and therefore how much they can be trusted. We have already outlined some of the problems associated with ranking evidence and methods, and for these reasons reviewers may choose not to make solid judgements on the strength of the evidence. However, at a minimum, it may be sensible to provide information on the kinds of (fieldwork and analytical) methods reviewed studies use, as well as on the features of the dataset their authors were working with. This way, the reviewers stop short of making judgements themselves, opting instead to let the reader draw their own conclusions about the strength of the evidence based on relevant information.

- **How to present analysis and findings.** When discussing the nature of a particular study or a wider evidence base, it can be helpful to present information on its characteristics visually – particularly for readers. One example of this is to use a 'traffic light' system to illustrate whether a reviewed study fulfils various criteria (e.g. provides extensive information on methodology, includes gender analysis, offers an explanation of possible causal mechanisms, and so on). Marking a study 'red' represents a failure to fulfil a particular criterion, 'amber' partial completion, and 'green' satisfactory completion. (Again, note that many of these assessments will not be bias-free.) When wanting to provide descriptive information on an evidence base more generally, the use of charts and graphs can be useful. As one example, in a forthcoming SLRC 'evidence brief' on the relationship between tax and livelihoods, the authors use a visual approach to illustrate some of the basic characteristics of the evidence base reviewed. From Figure 2 below, the reader can obtain an immediate sense of the thematic composition of the evidence reviewed, the kinds of methods used within each thematic category, and the extent of the evidence within each category.

Figure 2: An example of how to visualise the nature of an evidence base – classifying studies on tax and livelihoods



Source: SLRC (forthcoming)

4 Conclusion

This paper has outlined an empirically informed approach to doing a rigorous, evidence-focused literature review – an approach we have been using to a produce a literature review that adheres to the core principles of ‘full’ or ‘official’ systematic reviews while allowing for a more flexible and user-friendly handling of retrieval and analysis methods. Importantly, doing a rigorous literature review should be seen as a means to an end – helping produce a robust and sensible answer to a focused research question – and not an end in itself. This explains why we have placed less emphasis on documenting steps in the process, and more on retrieving, using and understanding the evidence.

So, what – specifically – is different about our approach? We argue that the process outlined in this paper differs from the standard systematic review method in seven clear ways.

1. Our process is not as strictly peer-reviewed as that used for systematic reviews. For example, one does not need to register the review and it is optional to have the protocol peer-reviewed.
2. Our approach does not require reviewers to collect as much information about the process, for example number of hits at different stages.
3. Our process requires reviewers to keep track of and collect information only on included studies – that is, not on excluded studies.
4. In the reviews we conducted, we often stopped short of assessing research quality and did not conduct a meta-analysis. Instead, we tried to use the findings in a sensible way to obtain a useful answer to the research question.
5. There is more flexibility within our approach to adapt the process if things do not work in practice or if practical challenges of various kinds are encountered. However, reviewers must account for alterations to the process, and transparency must be maintained throughout.
6. Our approach is more sensitive to the realities of the ‘information architecture’ found within the field of international development. It places a greater emphasis on locating grey literature and resources not found within the standard peer review channels, which is considered particularly important when a focus on impact evaluation is required.
7. We fully acknowledge the subjectivity inherent in different steps of the process and use it to our advantage, for example by using snowballing to include influential studies in the field.

We encourage experimentation with the process outlined here. Not every review serves the same objective, and retrieval, assessment and synthesis methods should be driven by the nature of the research question – not vice versa. Further, because different subject areas may be characterised by very different bodies of evidence – particularly in terms of methodologies and data – a degree of reflexivity and innovation may benefit the analysis phase. Because reviewers will not know (or at least should not know) what the evidence base looks like until after they have completed the retrieval and screening phases of their review, sticking rigidly to a predetermined assessment and analysis framework may not make sense in many cases.

Finally, a caution about the nature of literature reviews in general. Reviews, particularly those concerned with policy- and practice-focused research questions, are premised on the notion that there is an answer to a given research question, which can be 'found' simply by tapping into the relevant 'body of literature' and uncovering the 'evidence base'. This represents a very objectivist way of approaching the literature review, implying that there is not so much an art as a (hard) science to the process. However, while the notion of a 'body of literature' or an 'evidence base' suggests something tangible, bounded and cohesive, in reality the information required to understand a problem may be dispersed, hard to access (even inaccessible) and fragmented. We have already talked about how the 'information architecture' found within the field of international development prevents reviewers from easily retrieving information, but these issues apply too to the actual substance or content of that information. The terms 'body' and 'base' imply something connected up, something whole. However, researchers may draw on a range of different theories – if, indeed, they do so at all – and often approach a common problem from very different epistemological perspectives. The point being, if the 'body' or 'base' is characterised by such internal heterogeneity and variation, should we even be thinking about literature reviews in these terms?

In short, literature review tasks are complex. It is not as straightforward as digging up a monolith of information and assigning quality grades to its various features. In reality, it is unlikely that reviewers will be able to retrieve all available and relevant evidence to answer a given research question. And, even if we could, would we know how to make sense of it all? At the core of the process outlined in this paper is the argument that reviewers should not straightjacket themselves from the outset; adhering to some basic scientific principles may help improve the quality of a review, but – in our eyes – so does creating the space for innovation, adaptation and reflexivity.

References

Bergh, G. (2012) 'Systematic Reviews in International Development'. Internal. London: ODI.

Civil Service (2012) 'How to Do a REA'. Accessed June 2013: http://www.civilservice.gov.uk/networks/gsr/resources-and-guidance/rapid-evidence-assessment/how-to-do-a-rea

DeMarrais, K. and Lapan, S.D. (2004) *Foundations for Research Methods of Inquiry in Education and the Social Sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.

DFID (Department for International Development) (2013a) 'Systematic Reviews in International Development'. Accessed September 2013: https://www.gov.uk/government/publications/systematic-reviews-in-international-development/systematic-reviews-in-international-development

DFID (2013b) 'Assessing the Strength of Evidence'. How-to-Note. London: DFID.

Hagen-Zanker, J. and Leon, C. (2013) 'What Do We Know about the Impact of Social Protection Programmes on the Decision to Migrate?' *Migration and Development* 2(1).

Hagen-Zanker, J., Duvendack, M., Mallett, R. and Slater, R. (2012) 'Making Systematic Reviews Work for International Development Research'. Briefing Paper 1. London: SLRC.

Mallett, R., Hagen-Zanker, J., Duvendack, M. and Slater, R. (2012) 'The Benefits and Challenges of Using Systematic Reviews in International Development Research'. *Journal of Development Effectiveness* 4(3): 445-455.

Newton, B. J., Rothlingova, Z., Gutteridge, R., LeMarchand, K. and Raphael, J.H. (2012) 'No Room for Reflexivity? Critical Reflections Following a Systematic Review of Qualitative Research'. *Journal of Health Psychology* 17(6): 866-885.

O'Mara-Eves, A., Brunton, G., McDaid, D., Kavanagh, J., Oliver, S. and Thomas J. (2013) 'Techniques for Identifying Cross-disciplinary and "Hard-to-detect" Evidence for Systematic Review'. *Research Synthesis Methods* Early View.

Petticrew, M. and Roberts, H. (2006) *Systematic Reviews in the Social Sciences*. Oxford: Blackwell Publishing.

Popay, J., Roberts, H., Snowden, A., Petticrew, M., Arai, L., Rodgers, M. and Britten, N. with Roen, K. and Duffy, S. (2006) 'Guidance on the Conduct of Narrative Synthesis in Systematic Reviews'. Accessed September 2013: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.178.3100&rep=rep1&type=pdf

Pritchett, L. and Sandefur, J. (2013) 'Context Matters for Size: Why External Validity Claims and Development Practice Don't Mix'. Working Paper 336. Washington, DC: CGD.

Shemilt, I., Simon, A., Hollands, G.J., Marteau, T.M., Ogilvie, D., O'Mara-Eves, A., Kelly, M.P. and Thomas, J. (2013) 'Pinpointing Needles in Giant Haystacks: Use of Text Mining to Reduce Impractical Screening Workload in Extremely Large Scoping Reviews'. *Research Synthesis Methods* Early View.

SLRC (Secure Livelihoods Research Consortium) (forthcoming) 'Taxation and Livelihoods: Evidence Brief'. London: SLRC.

Thompson, J., Davis, J. and Mazerolle, L. (2013) 'A Systematic Method for Search Term Selection in Systematic Reviews'. *Research Synthesis Methods* Early View.

van der Knaap, L.M., Leeuw, F.L., Bogaerts, S. and Nijssen, L.T.J. (2008) 'Combining Campbell Standard and the Realist Evaluation Approach: The Best of Two Worlds?' *American Journal of Evaluation* 29(1): 48-57.

Waddington, H., White, H., Snilstveit, B., Garcia Hombrados, J., Vojtkova, M., Davies, P., Bhavsar, A., Eyers, J., Perez Koehlmoos, T., Petticrew, M., Valentine, J.C. and Tugwell, P. (2012) 'How to Do a Good Systematic Review of Effects in International Development: A Tool Kit'. *Journal of Development Effectiveness* 4(3): 359-387.

Walker, D., Bergh, G., Page, E. and Duvendack, M. (2013) 'Adapting a Systematic Review for Social Research in International Development: A Case Study from the Child Protection Sector'. London: ODI.

ODI is the UK's leading independent think tank on international development and humanitarian issues.

Our mission is to inspire and inform policy and practice which lead to the reduction of poverty, the alleviation of suffering and the achievement of sustainable livelihoods.

We do this by locking together high-quality applied research, practical policy advice and policy-focused dissemination and debate.

We work with partners in the public and private sectors, in both developing and developed countries.

ISSN (online): 1759-2917

ISSN (print): 1759-2909

Cover image: Jan Moren (CC-BY-NC)

Overseas Development Institute
203 Blackfriars Road
London SE1 8NJ
Tel +44 (0)20 7922 0300
Fax +44 (0)20 7922 0399

Secure
Livelihoods
Research
Consortium