Gaps in Evaluation Methods for Addressing Challenging Contexts in Development

Calum Davey[1], Syreen Hassan[1], Chris Bonell[1], Nancy Cartwright[2], Macartan Humphreys[3], Audrey Prost[1], James Hargreaves[1]

[1] London School of Hygiene and Tropical Medicine
[2] Durham University and University California, San Diego
[3] University of Colombia

Colophon

Suggested Citation: Davey C, Hassan S, Bonell C, Cartwright N, Humphreys M, Prost A, Hargreaves J, 2017 Gaps in Evaluation Methods for Addressing Challenging Contexts in Development. CEDIL Pre-Inception Paper: London

About CEDIL: The Centre of Excellence for Development Impact and Learning (CEDIL) is an academic consortium initiative supported by UKAID through DFID. The objective of the centre is to commission and implement impact evaluations, promote the uptake and use of evidence from impact evaluations, and develop and demonstrate new and innovative methodologies for impact evaluation and evidence accumulation. Centre for Evaluation, London School of Hygiene and Tropical Medicine is one of a number of CEDIL members that form its intellectual leadership team.

Corresponding Author: Mr Calum Davey, email: calum.davey@lshtm.ac.uk

Table of Contents

Abstract

We currently do not learn as much as we could from evaluations. While there are well-established methods for determining, and understanding, the effects of simple interventions in one set of places, it is less clear how to learn the most possible from evaluations of context-specific, complex interventions, and apply what we learn to other contexts. This is especially important in international development: evaluations are limited by time, cost and opportunity, and there is substantial heterogeneity in the issues and contexts within which work is undertaken. We consulted with evaluation experts in the Centre for Evaluation at LSHTM and an interdisciplinary group from the Intellectual Leadership Team at CEDIL.

The consultations identified gaps between established methods and their use, and gaps in available methods. Gaps in use included approaches to interpreting sub-group analyses, identifying unintended effects, and using theories of change. Gaps in methods included choices of non-randomised designs, and how to conduct mediation analysis. However, the pre-eminent gap identified was how best to use findings from evaluations to inform policy decisions. We argue that effects of complex interventions are often highly context-dependent and that the theories that describe interventions in context are what could be transported from one place to another to maximise learning from evaluations. Methods are lacking for how to develop and test such theories, and how to know that a theory developed and strengthened in one place and time is appropriate to inform action elsewhere.

Section 1

Background

1.1 What is Evaluation?

Evaluations vary in terms of the questions that are asked, the methods used, and the overall aim of conducting the evaluation (Rossi et al., 2003). Evaluations of interventions in most disciplines implicitly or explicitly follow one of two forms, depending on the overall aim and the research question they try to answer, often referred to as 'impact' and 'process' evaluations (Moore Graham F et al., 2015). While these activities are in principle defined by the questions that they are asking, in practice there is a tendency to apply quantitative methods in impact evaluation, while a mix of quantitative and qualitative methods are generally applied in process evaluation.

'Impact' evaluations aim to answer the question: 'what effect did the intervention have?' They often estimate causal effects using counterfactual analysis of quantitative data reporting on the intended outcomes, e.g. observational quantitative analysis, experimental designs, and quasi-experimental designs (Gertler et al., 2011). Methods that draw on qualitative data are also used, such as qualitative comparative analysis (Ragin, 1999). Some impact evaluations will not be able to estimate the counterfactual using data from a credible comparison group, although the counterfactual logic is often still applied (Mohr, 1995). In before-after, small-N (White and Phillips, 2012), and theory-based evaluations (Weiss, 1997), what happened when the intervention was implemented will be compared with an estimate, variously generated, imagined, and expressed, of what would have happened had the intervention not been implemented – i.e. the counterfactual.

Impact evaluations estimate overall effects on intended outcomes *in situ*: in the context(s), including the times, when and where the intervention was implemented. They are often done by comparative studies that aim to minimise bias by estimating the average association between intervention exposure and intended outcomes, while minimising confounding, reverse causality, selection and information bias (Gertler et al., 2011). The overall effects (or lack of) may mask subgroup heterogeneity (Wang, 2007), and many impact evaluations will also estimate sub-group effects *in situ*. Some impact evaluations may also estimate effects on outcomes that are not intended, and explore potential unexpected benefits, harmful externalities, or paradoxical effects (Bonell and J, 2014).

'Process' evaluations aim to measure and understand processes *in situ*, balancing breadth and depth. Process evaluations describe processes of planning, implementation and receipt, analyse the role of contextual factors, and explore mechanisms of action (Moore Graham F et al., 2015).

Process evaluations use descriptive data to provide post hoc explanation of impact findings (e.g. whether a null result is explained by poor fidelity). Process evaluations are used to build and refine theory about contextual factors affecting implementation; hypotheses are tested in new analyses, often drawing on process and impact data (e.g. did baseline local infrastructure or client profile moderate intervention effects or did improved provider skills mediate intervention effects?). Process evaluations examine mechanisms. They may attempt to uncover whether effects on distal outcomes can be explained by effects on process or proximal outcome measures.

In describing the aims of impact and process evaluations above, we have stressed that the immediate objective is usually to answer the evaluation questions as reliably as possible *in situ*. This is sometimes referred to as the 'internal validity' of the findings (McDermott, 2011). Of equal importance is the task of using the evaluation to efficiently predict impacts and processes in other settings, and thus guide decisions. Evaluations rarely assume that the effects and processes observed *in situ* would be the exactly the same if the intervention was delivered in a different context or different time. It has been argued that we should assume that the results cannot be generalised as a starting point (Cartwright, 2011), that they have low 'external validity'. Exploring the implications of existing evidence for future actions is a key challenge for evaluations.

Impact and process evaluations together aim to estimate the effects of the intervention, and describe the processes that afforded those effects. Evaluations often also include questions about costs and sustainability in economic evaluations, which are combined with the estimated effects and processes to inform policy decisions for elsewhere. The Development Assistance Committee of the Organisation for Economic Co-operation and Development recommends that evaluations determine relevance, effectiveness, efficiency, impact, and sustainability, to inform policy decisions (Chianca, 2008) (in public health, the RE-AIM framework is similar (Glasgow, 1999)). Neither an impact or process evaluation could fully answer any one of these on its own; they are best answered by the combined analysis and interpretation of both, with additional data from economic evaluation.

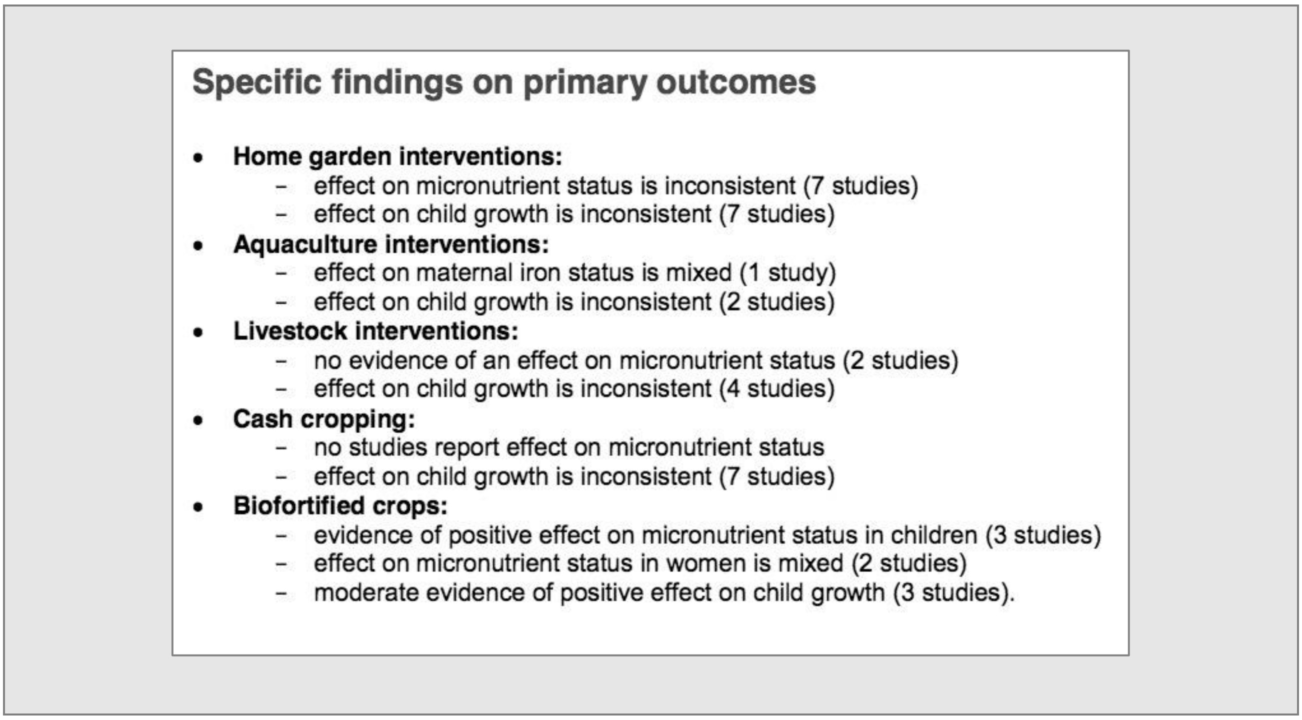
1.2 What is Evaluation For?

Evaluations are conducted for different purposes. Some evaluations are for deciding whether to continue implementation of an intervention or how to scale-up and refine the intervention in the context where it is conducted. (It is worth noting that contexts are never static, and critical elements – such as the political climate or institutional strength – may vary over time as much as over space or between populations (Rosenzweig and Udry, 2016).) Others are conducted to help determine where or for whom it is appropriate to transfer the intervention to, or how to adapt the intervention for other settings or populations. Formative research assesses context to determine if an intervention will likely be appropriate and effective in a new setting.

Achieving these aims is a challenge for evaluation in the field of international development. Some problems are relatively case-specific, e.g. there are many differences between each disaster or conflict. Other aspects of human development, such as education or agricultural income, are so widely experienced that there is considerable contextual variation across the globe. Resources are limited for evaluation in the sector. Therefore, it is important to maximise learning.

This report describes 'gaps' in the methods and practice for evaluating interventions to inform policy decisions in development. To illustrate the challenge, **Figure 1** reproduces the summary of the evidence for effects of agriculture interventions on nutrition outcomes from a 2014 DFID evidence paper (DFID, 2014). Many of the results are inconsistent or mixed; therefore, policy recommendations are difficult to make despite several evaluations for each type of intervention. Our hypothesis is that there are ways of conducting these evaluations and synthesising the evidence from them that can give us more useful information.

Figure 1: Summary of the evidence for the effect of agricultural interventions for nutritional outcomes from 2014 DFID evidence report. Key – Inconsistent: similar studies with findings that do not concur. Mixed: dissimilar studies on the same topic that do not concur, or a variable pattern of impacts within the same study (p. 7).



Specific findings on primary outcomes

- **Home garden interventions:**
  - effect on micronutrient status is inconsistent (7 studies)
  - effect on child growth is inconsistent (7 studies)
- **Aquaculture interventions:**
  - effect on maternal iron status is mixed (1 study)
  - effect on child growth is inconsistent (2 studies)
- **Livestock interventions:**
  - no evidence of an effect on micronutrient status (2 studies)
  - effect on child growth is inconsistent (4 studies)
- **Cash cropping:**
  - no studies report effect on micronutrient status
  - effect on child growth is inconsistent (7 studies)
- **Biofortified crops:**
  - evidence of positive effect on micronutrient status in children (3 studies)
  - effect on micronutrient status in women is mixed (2 studies)
  - moderate evidence of positive effect on child growth (3 studies).

Section 2

Methods

We conducted an interdisciplinary consultation and literature search to identify 'gaps' in methods used to evaluate interventions in international development. We differentiated between gaps in the adoption of existing methods and best practices, and areas where methodological research is needed.

2.1 Consultations with Centre for Evaluation (CfE) and CEDIL Intellectual Leadership Team (ILT)

We invited members of the CfE at LSHTM to attend a lunch-time consultation on the gaps in evaluation methods. On the 21 March 2017, we held a full-day consultation and workshop with three non-LSHTM members of the CEDIL ILT who represent fields outside of public health. They were asked to give 45-minute public presentations on gaps in evaluation methods. The afternoon was spent discussing, sharing literature, and constructing a concept map.

2.2 Key Literature

Following-up on issues raised in consultations, we explored literature relevant to the 'gaps' that were consistently identified and emphasised by the CfE and ILT consultations.

2.3 Report and Presentation

A draft report was shared with the CEDIL management team and DFID, and presented at a CEDIL workshop attended by all members of the ILT on the 25th of April 2017. We received feedback, including from CEDIL Research Director Howard White and two peer reviewers. A second draft was submitted to CEDIL and shared with DFID. Further comments from DFID were received on August 29th 2017.

Section 3

Results

The consultations with the CfE and the ILT identified many 'gaps' in evaluation methodology. Loosely, the gaps can be grouped into two types. 'Gaps in use' are gaps between expert consensus and use of the methods. Methodological innovation is not needed to close these gaps; the challenge is awareness and implementation. 'Gaps in methods' are where innovation is needed to improve learning from evaluations. This report is focused on the latter gaps, but we first briefly describe the first set of gaps.

3.1 Gaps in Use

There is little disagreement among experts or in the literature that the randomised controlled trial (RCT) produces the most internally valid estimates of *in situ* intervention effects; the gap is in the application of RCTs where possible. This is not a methodological gap: RCT methodology is a crowded market, with contributions from the fields of medicine, economics, psychology, statistics, education, criminology and agriculture. Dissent against RCTs is largely on philosophical grounds (Marchal, 2013), incorrectly interpreting the RCT design as an epistemological construct, rather than simply as a way of creating (partially) understood and controllable variation in intervention exposure (Bonell, 2012). Opposition to the use of RCTs in some contexts or for certain kinds of intervention has been made on the grounds of practicality, although often this is unfounded (Rosen, 2006), and there is evidence of a gap in applying RCTs when they would be applicable and useful. RCTs can be adapted while retaining their advantages over alternative designs (Craig, 2008), and have been used in challenging contexts (Blanchet, 2017).

There is also a broad consensus on the use of subgroup analysis to estimate heterogeneity of effects *in situ*, but a gap in the use of sensitive methods that avoid data dredging. 'Eyeballing' the effects for different subgroups is insufficient, and formal statistical methods should be used. However, evidence for heterogeneity of effects often lacks precision because sample sizes are determined for the precision of the overall intervention effect (Wang, 2007). To improve the effective power of the tests after accounting for multiple testing (Shaffer, 1995), tests should be based on *a priori* hypotheses and limited in number.

There is a consensus on the need to examine unintended effects, but a gap in the use of appropriate theory-led, mixed-methods approaches. The means to examine unintended effects are relatively uncontroversial: evidence-based hypotheses should be used to inform additional data collection for key unintended outcomes (Bonell and J, 2014), and statistical analyses will need to be focussed to avoid data dredging. Unexpected outcomes cannot be anticipated, and qualitative, open-ended research methods can capture these and inform further data collection in the same or future evaluations. Evidence of an increase in rare adverse events should be explored (e.g. catastrophic side effects).

All interventions presuppose theory even if implicitly. There is a consensus on the need for a 'theory of change', logic model, or another expression of the theory behind why and how the intervention is supposed to affect outcomes (Moore Graham F et al., 2015). Theories of change take a variety of forms: linear and non-linear, simple or complex (Davies, 2004). Superficially, there is a gap when theories of change are not articulated and agreed by the teams implementing and evaluating the interventions, and therefore are not available to guide the evaluation or to express a shared set of assumptions about how the intervention might affect outcomes. Substantively, and more commonly, there is a gap when the theory of change does not reference and reflect existing scientific (or implementation or programme) theory and is not grounded in the context where the evaluation is taking place, drawing on local stakeholders and formative research. There is a gap when the implementation theory is not explicit; the theory of change should describe how the intervention will be delivered within real-world systems, as well as the theorised mechanisms behind the active ingredients (Breuer, 2016). There is also a gap in that even when the theories of change specify mechanisms, they do not specify contingencies, e.g. the contexts, which determine whether those mechanisms are triggered or generate outcomes (Popper, 2002) (Pawson and Tilley, 1997).

There is a gap in the application of best practices for process evaluation, in particular, the use of mixed-methods. Process evaluations should include quantitative indicators of fidelity, reach, acceptability; and qualitative exploration of a variety of accounts of conditions for, and meanings and consequences of actions (Moore Graham F et al., 2015). There is a consensus on a need for breadth through quantitative data and depth through qualitative data.

The consensus that these methods are usually appropriate and helpful for evaluations does not mean that they are always considered or used. This 'gap' is a gap between theory and practice: dissemination and translation activities could help to close this gap.

3.2 Gaps in Methods

Although methods are available, there are aspects of evaluation design that are contentious, experts disagree on the best approaches, and further work is required to refine the design decisions facing evaluators.

There is a gap in the reliability of the interpretation of intervention effects *in situ* when using non-randomised methods. There is an incomplete consensus regarding how to best estimate effects when an RCT is inapplicable. Non-randomised methods often work by exploiting variation in the delivery of the intervention and the available data (Davey et al., 2014); therefore, the strength or weakness of a design will depend on the particular circumstances, and the risk of bias will be appraised with methods specific to the design that was used. Furthermore, there is variation between disciplines in how and when methods are applied and how their results are reported. All things being equal, controlled-before-and-after designs, i.e. comparison with a control group while accounting for baseline differences arising from the lack of random allocation, are often the most robust option when possible (Bonell, 2008) (Shadish et al., 2002). For example, difference-in-difference designs, and the recently developed synthetic control method that makes use of repeat collection of data from the same places (although not always the same people) (Abadie, 2010). Interrupted-time-series designs can be used when there is no control group, but require measures of the main outcomes at many time points to develop a model of the trend in the outcome (Bernal, 2017). The least reliable design uses only cross-sectional data to estimate the association between exposure to the intervention and outcomes at one-time point. However, in some special circumstances, regression discontinuity analysis can be used to get high-quality estimates of causal effects with cross-sectional data only (Hahn, 2001).

There is a gap in the understanding of mechanisms. Methods for measuring and understanding intervention mechanisms *in situ* are also not widely agreed upon. Qualitative data collection can be useful but is often collected from a small sub-sample and limited to mechanisms that are revealed, in some way, to the participants in the interviews or focus groups. Process evaluations that are limited to interviews and focus groups may not properly interrogate mechanisms because participants may rely on narrative 'common-sense' explanations for the changes that they think they have observed (which themselves may be confounded by secular trends). Naturalistic observational methods, such as observation or repeat interviews, allow the researcher to 'observe' and not rely only on what that small group of participants say; it is possible that it is not the qualitative approach that is limited, but the choice of the methods, due to time and cost. What can be learned about mechanisms from qualitative approaches may require more attention to sampling, and ongoing data collection throughout implementation.

Quantitative mediation analysis can be used to identify effects across larger samples. The Baron and Kenny method to estimate mediators has limitations, especially for binary outcomes (VanderWeele, 2010), and confounding needs to be accounted for even when the intervention is randomised (Richiardi, 2013). Different fields have developed new ways of approaching mediation (Daniel, 2011) but these have not been applied to evaluations of complex interventions in challenging contexts. These new methods will need to be appraised, with a focus on their ease of implementation, need for dedicated experts, and data requirements.

There is a gap in the methods for positioning process and impact evaluations relative to each other, and drawing on them jointly to answer evaluation questions. To explore the extent of this gap and get a sense of the principles that we would need to consider when bringing together impact and process data, we identified frameworks from other methodological areas that have similar aims: mixed-methods research, Bayesian integration of qualitative and quantitative data, and evidence synthesis for complex interventions.

Mixed-methods research literature describes three strategies by which different kinds of data (in this case quantitative and qualitative) can be combined: embedding, connecting, and merging (Creswell, 2011). All three strategies are relevant to combining process and impact evaluations and help us explore this gap.

Embedding is when research of secondary priority is set within a primary design. Often, process evaluations are embedded in the ‘main’ impact evaluation. In a review of mixed-methods designs in mental-health-services research, Palinkas *et al.* gave an example of embedding (Palinkas, 2011):

> This type of mixing was more commonly found in randomized controlled trials and in implementation studies, where qualitative studies of treatment or implementation process or context were embedded within larger quantitative studies of treatment or implementation outcome. For instance, to better understand the essential components of the patient-provider relationship in a public health setting, Sajatovic and colleagues conducted a qualitative investigation of patients’ attitudes toward a collaborative care model and how individuals with bipolar disorder perceive treatment adherence within the context of a randomized controlled trial evaluating a collaborative practice model. (p. 260)

Embedding in this way may lead to greater emphasis on answers addressed by impact evaluations, such as average intervention effects *in situ*, than from process evaluations, such as the appropriateness of the intervention and the mechanisms of effect. Process evaluations can be seen as adaptable and having the auxiliary function to ‘help explain’ impact evaluation results, but not the other way around – the impact evaluation design is ‘fixed’ and is often used to determine the size of the evaluation. This may not be defensible on the grounds of what are the most important questions for policy, and the notion of embedding reveals a gap in how process and impact evaluations are emphasised.

Connecting involves using information from the analysis of one set of research to develop or inform subsequent data collection or analysis in another. Creswell *et al*. offer the example (Creswell, 2011):

> Dawson et al. studied non-abusing drinkers diagnosed with hepatitis C … An initial qualitative component based on interviews and Internet postings described new decision factors related to curtailing the consumption of alcohol. These findings were used to develop new items for a quantitative instrument, which was administered in the second phase to assess the prevalence of the new factors and their association with current drinking. (p. 5)

The timing of connecting data in process and impact evaluations require guidance since analyses can be lengthy (Aarons, 2012). Transparency is also a consideration when changes are made to data collection protocols or outcome measures during the evaluation and in response to observations of intervention delivery. There is a gap in the guidance available for evaluators connecting process and impact aspects of their evaluations.

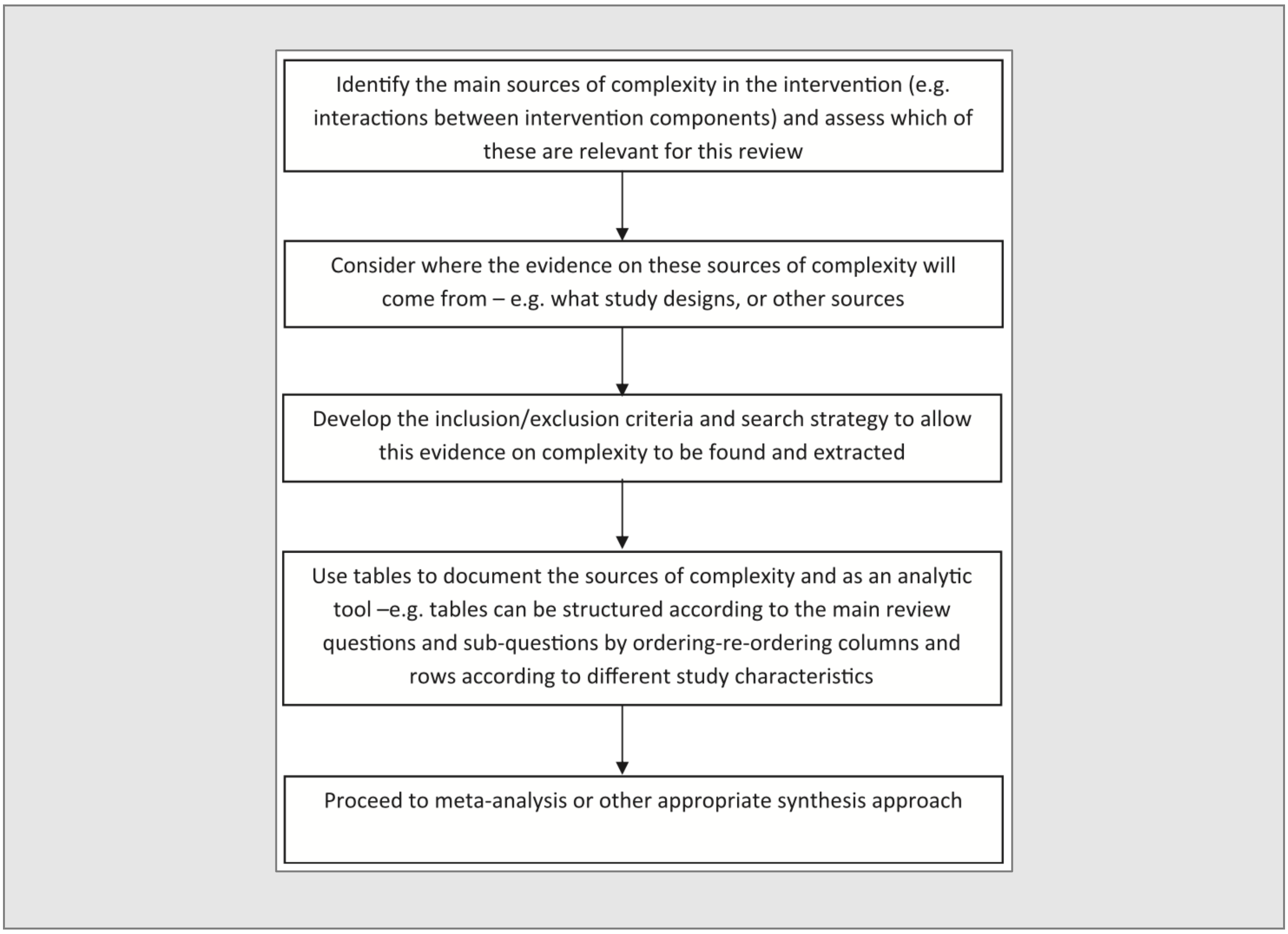
Merging combines different types of data to draw conclusions (Carvalho and White, 1997). Merging can be done by triangulation – where one type of data is used to confirm the conclusions obtained from the other type of data – or by transforming the results of one type of data to the other type and then analysing them together, e.g. by quantifying qualitative data and using statistical methods. Impact and process evaluations are interpreted together when considering the generalisability of the findings to other contexts (Burchett, 2013), for identifying and estimating sub-group effects, and understanding mechanisms (Oakley, 2006). However, evaluators find it difficult to merge impact and process evaluations since there are no widely used models or processes for interpreting multiple data sources jointly. Instead, the process of bringing together various forms of an evaluation is often informal; rarely refers to any framework or methodology; the methods, including any iteration between theory and data, are not pre-specified; and the results of the joint interpretation are often described *ad hoc* in the discussion section of evaluation reports and academic papers.

This lack of methodology and frameworks for merging process and impact evaluations is a major gap in evaluation methods. Bayesian statistical approaches show promise for combining subjective and qualitative information with quantitative data from different sources. A Bayesian method that has been developed by Macartan Humphreys with Alan Jacobs combines impact evaluation data with data from the process evaluation (Humphreys, 2015). The method requires theory in the analysis, which makes explicit the strength of prior beliefs about the strength and structure of the causal connections between elements of the theory. In their paper, Humphreys and Jacobs apply the method to two examples from political science. One asks whether, and by how much, a country having natural resources (the treatment, or ‘intervention’) causes conflict (the outcome). The analysis uses qualitative process-tracing data collected from countries that engaged in civil war between 1990 and 2000, and data from countries with and without resources, and with or without conflict, during the same period. The qualitative and quantitative data were linked through the parameters of a Bayesian probability model that described the likelihood of there being a causal effect given the data observed. The data were transformed into quantitative measures, which were then merged to estimate the effect sizes *in situ* using both sources of data, as well as simultaneously updating beliefs in the causal connections implied by the theory. By using a mathematical model, the approach is pre-specifiable and transparent. This means that any biases, in the form of prior beliefs, are explicitly stated, and the effect they have on the conclusions can be assessed using sensitivity analyses; in contrast, informal approaches to combine data types will have implicit biases that are not reported. The method requires quantification of theorised causal-relationships, which might be too much to ask of evaluators. Examples of this kind of approach in practice are few (Befani and Stedman-Bryce, 2017), it requires trained researchers to apply it, and having well-coordinated multi-faceted research teams to generate the quantified qualitative data and agree on the causal-model assumptions. Research to retain the strengths of the method while making it accessible to practitioners is on-going (Befani et al., 2016).

While the lack of a framework to combine data from process and impact evaluations is the major gap in methods, there is an auxiliary gap in defining a transparent and pre-specifiable process for identifying, appraising, and weighting the data from across an evaluation. Pre-specification is regarded as a safeguard against conscious or unconscious bias in the analysis of data or results, even when researchers choose for good reasons to deviate from their pre-specified approach (Chan et al., 2013). In practice, for all interventions and contexts some of the analysis needs to be pre-specified to avoid bias (i.e. the primary and secondary outcomes) and some of it needs to be iterative to ensure that we draw on emergent, unexpected findings from open-ended qualitative research to focus our analyses of quantitative data on the most plausible and sophisticated hypotheses possible. Reviews of the application of mixed-methods approaches found that the rationale for mixed-methods research is often unspecified, or that what was achieved, was different from what was planned (Palinkas, 2011) (Bryman, 2006). Mixed-methods research may offer hidden ‘added value’ that goes beyond what researchers foresee (Bryman, 2006); similarly, the integration of an evaluation may present an issue. Thus, a key challenge is determining for which questions pre-specification is most useful, and for which iterative or 'grounded theoretical' approaches are more appropriate, allowing the integration of results to guide and shape our understanding in answering the evaluation questions and the methods used to do so. Although iteration between theory and data may be appropriate, we need to be cautious about using methods which are prone to explicit or implicit bias, and data mining (White, 2002).

Literature on systematic reviews of complex interventions may be useful for developing guidance for pre-specifying the joint interpretation of impact and process data from within an evaluation. Recommendations such as defining the process for gathering data sources, inclusion criteria, quality assessment, and extraction methods in advance are relevant for merging multiple data sources: see figure 2, from Petticrew *et al*. (Petticrew, 2013).

Figure 2: Initial steps in identifying and analysing complexity (Petticrew et al., p 1235).



Identify the main sources of complexity in the intervention (e.g. interactions between intervention components) and assess which of these are relevant for this review

Consider where the evidence on these sources of complexity will come from – e.g. what study designs, or other sources

Develop the inclusion/exclusion criteria and search strategy to allow this evidence on complexity to be found and extracted

Use tables to document the sources of complexity and as an analytic tool –e.g. tables can be structured according to the main review questions and sub-questions by ordering-re-ordering columns and rows according to different study characteristics

Proceed to meta-analysis or other appropriate synthesis approach

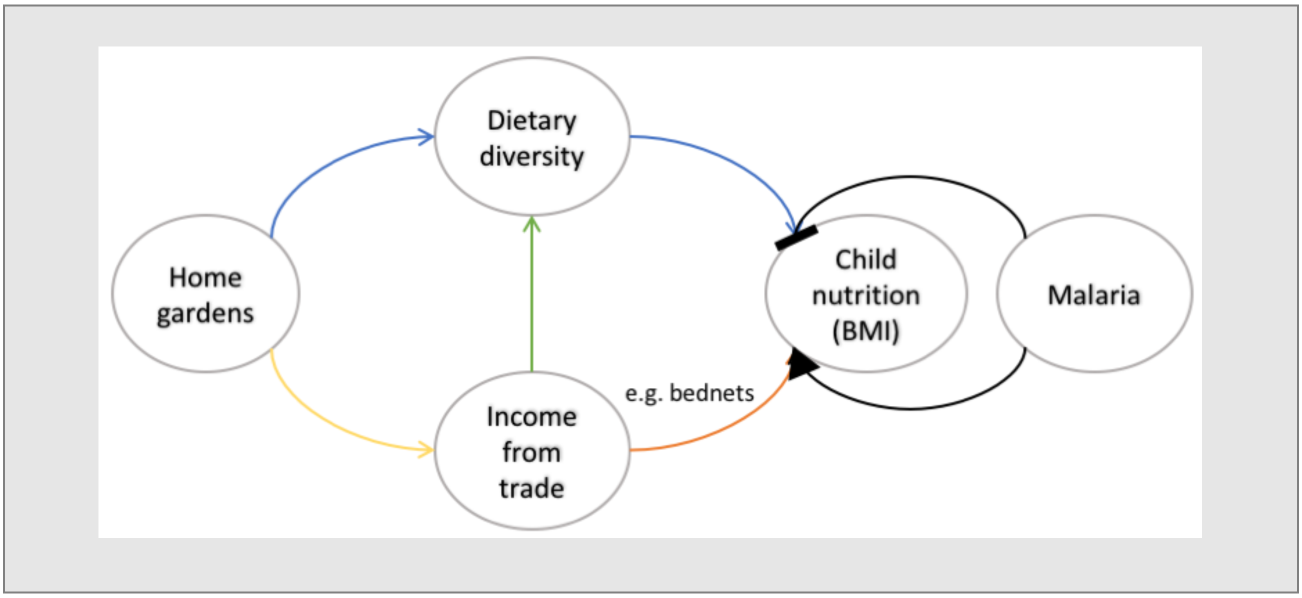
In summary of the methods gaps identified so far: there are gaps in reliably identifying *in situ* causal effects when randomisation is not possible. There are gaps in understanding mechanisms because of the limitations of qualitative methods and the difficulty of using valid quantitative methods in practice. And while most evaluations involve process and impact elements, there is a gap in the guidance on best practices for using them together to maximise learning. Process evaluations are often uncritically embedded within impact evaluations, and there is limited guidance on how to connect data sources in a timely and transparent fashion. Process and impact evaluations can, and often are, merged to make the final inferences about the key evaluation questions, but there is a gap in the methods available to guide this process, and ensure that inference is transparent and replicable.

The gaps identified above relate to estimating causal effects and understanding causal mechanisms. Both are essential if evaluations are to provide learning from one setting that can be usefully used to inform decisions in other settings. A further gap relates to the optimal methods and approaches to understanding the role of context and characterising the context(s) within which an evaluation is undertaken in such a way that allows findings to inform robust policy recommendations in other contexts and at other times.

ILT member Nancy Cartwright contrasted intervention-centred and context-centred approaches to evaluations. Intervention-centred approaches take the view that the intervention is goal-directed towards the outcome. We think about a lot of simple interventions this way, such as drugs. So long as there is a need for the medication, we have a sense of an 'intervention-outcome' pairing between the drug and the outcome: the challenge for the evaluator (in this instance, a clinical trialist) is to work out the magnitude of the causal effect. The context-centred approach views the effect of an intervention as that which is afforded by the context. We tend to think about complex social interventions in this way, where the effect of an intervention on health and development outcomes depends on everything else in the system and how it is all related. However, simple interventions also seem more context-centred when you look closely: even the effects of drugs are dependent on myriad features of the bodies they are in (Hey, 2015).

We describe a hypothetical evaluation of gardens to improve child nutrition as an example of the challenge of using evaluation findings from context-centred interventions, linking back to the illustration of the problem in the Background. This is a seemingly simple intervention; however, it illustrates the problem that would most likely be even greater with a complex multi-level program. A theory of change is shown in Figure 3**,** adapted from the DFID evidence report modified from (Sharma, 1999: 20). The theory links the specific intervention, home gardens, to one aspect of the context – malaria prevalence – and childhood BMI. In this theory, gardens are thought to improve nutrition by increasing dietary diversity, either directly (blue path) or via increased income from selling produce (yellow-green-blue path), and/or by increasing access to non-food technologies such as bed nets (yellow-orange path). Note that this is an example of a relatively linear theory of change, omitting various potentially important contextual factors such as existing household and community resources, food preferences, seasonality, background food security, and market-related factors that influence households' ability to generate income from the produce of their home gardens. These factors are omitted to simplify the illustration.

Figure 3: Simplified diagram adapted from DFID evidence report on agriculture and nutrition.



In the presence of malaria, increasing dietary diversity is not sufficient to improve nutrition status, i.e. the blue pathway is 'blocked' (shown with a black bar). In the theory, higher malaria burden will enhance the effect of increased income on BMI because families can use the money to purchase bed nets, a highly efficacious intervention to improve infant health (shown with a black triangle). The prevalence of malaria will vary between contexts. The effect observed in an evaluation will depend on the mechanism of action and on the prevalence of malaria. The gardening intervention will have a different effect on child mortality in a context where: (1) treated bed nets are free and available (2) bed nets are sold for a small price (3) bed nets are not for sale locally. Plus, whatever difference the availability of bed nets makes will depend on the prevalence of malaria (i.e. the burden of disease), and malaria treatment. Therefore, the effect estimate (or estimates, from sub-groups) on their own do not provide good predictions for the effects of policies elsewhere.

Rather than only focusing on the effects of the intervention, the evaluation can be used to test the theory. Strengthening theory may help with transporting findings between contexts. For example: if the impact evaluation finds a null result (i.e. no effect of home gardens on BMI) with a high malaria burden, then the income from trade should also be unchanged by the intervention. By investigating the moderating (e.g. malaria prevalence) and mediating (e.g. income from trade) 'support factors' we can perhaps test and refine the theory. Some evidence will be inconsistent with the theory, for example, if we found a null overall impact when household income was increased by the intervention. Perhaps bed nets are free (and therefore not affected by income changes) or not used to cover children at night; these additional factors would be added and incorporated into the theory. We do not, however, have clear methods for knowing the limits of the theory, or whether the theory structure holds in other contexts (see below).

ILT member Audrey Prost shared an example of how a theory-centred approach to evaluation has been applied to work from India on the impacts of participatory interventions with women's groups (Manandhar, 2004). The research included randomised controlled trials as well as evaluation of large-scale 'scaling-up' interventions with a controlled, prospective, non-randomised quasi-experimental design (Prost, 2017). A deliberative-democratic process was used to develop an explicit theory describing the intervention and the context. The term 'deliberative democracy', coined in 1980 by Joseph Bessette (Bessette, 1980), refers to a law-making process in which laws are deliberated with citizens who can express their opinions and influence the process. In this context, theory was developed using a process in which stakeholders, implementers, evaluators, and people from the population intended to receive the intervention discussed and made decisions jointly. However, Audrey reflected that many decisions about this process were taken *ad hoc* and that guidance to help evaluation teams develop theory, who should be involved, and what form the theory should take would be highly welcome. While there is now more guidance on collaborative development of theories (Moore Graham F et al., 2015) (Silva, 2014), the challenge remains that the people involved in the discussion are often immediate implementers and researchers, so the systems element of context are under-documented. The gap is examples (rather than generic guidance) on who to engage in theory development so that issues of transferability are addressed.

Recognising that the context matters, and developing theory, leads to challenges. Understanding and specifying everything about the system is difficult: we cannot map the whole system. While there are ways to present the overall organisation of a theory of change using box-and-arrow diagrams, there is a gap in how to define and represent the causal functions that link elements together. This is because causal relationships between just two variables can take different forms (e.g. linear, multiplicative), and between many variables there can be complex interactions (e.g. one factor may need to be present for another to have an effect). These causal effects are afforded by the context, and it is difficult to determine the number of contextual elements to include in the theory and the structure of how they relate. While it is impossible for research to attend to all aspects of context, perhaps evaluators could choose to focus on aspects that are implicated in the initial theory (Popper, 2002). Qualitative research could be used to get glimpses into other aspects; however, even open-ended qualitative research probably cannot attend to everything and is rooted in prior theory, formally or informally. It may be appropriate to remain highly sceptical of the guidance that the theory is offering, and narratively investigate 'exceptional responders' or 'positive deviants' to generate initial theory (Roche, 2015). We may need to accept a pluralistic approach, with different researchers using different theories; we may learn more by being more deliberate in ensuring there is a good diversity of teams looking at different theories, rather than following a particular fashion (e.g. social marketing last decade; nudging this decade). Questions arise, such as: who should be involved in developing the theory? What form should the theory take? How should the underlying structures be represented and communicated? These are on-going areas of research that are not easily resolved.

Once the important parts of the system are understood, the various mediators and moderators could, one hopes, be measured and used to predict the effects of the intervention in a new context with different levels of these factors. To return to our simplified example, if we are able to study how differently our intervention works in settings with different prevalences of malaria, we could predict the effect in a new place with a new malaria prevalence. However, the usefulness of understanding the role of contextual elements in one set of places for predicting the effects elsewhere relies on the causal functions that relate the factors to one another being the same – e.g. the relationship between malaria, nutrition, and gardens. Fundamentally, we do not have good methods for quickly identifying if the structure of our theories in one context applies in another. In our example from India, the theory about the effectiveness of women's groups to improve maternal and new-born health was refined throughout a series of related evaluations. However, it was unclear which other contexts it could appropriately be used in, or how much the theory can be changed for a new context, such that the policy predictions remain reliable. Some work has been done in the causal analysis literature on how similar the structures in two contexts need to be for effects to travel (Pearl, 2011): work in this area could improve the usefulness of evaluation findings by giving guidance on where and when theories from one evaluation setting can be used for policy recommendations, and what key pieces of data would be most informative.

We have argued that the use of findings from an evaluation to inform policy requires knowledge of the applicability and compatibility of the theories for the underlying structures in the original and target contexts. This is the big gap in evaluation methods that requires methodological development.

Section 4

Next Steps

We identified gaps in the methods for evaluating interventions for international development. Translating findings from one context to another is the key challenge. We argue that an approach which uses theory is appropriate for complex interventions, however, there are substantial gaps in methods for its application in practice.

Our approach to identifying gaps was rapid and expert-led and therefore limited by the participants in the consolations, and the literature that we were directed towards. Although the ILT consultation was with a multidisciplinary group, the main authors at the CfE at LSHTM are mainly aware of conventions within public health. We have focused on gaps in methods for answering the evaluation questions about effects and mechanisms associated with interventions and outcomes in certain contexts, and how can these be useful for policy. We have not considered gaps related to costs and economic evaluation, or methods to reducing the time required to generate and communicate evidence.

CEDIL should seek to identify and appraise potential ways of better generating and using findings from evaluations in international development to inform policy. Developing methods for combining impact, process, and theory offers great promise to help DFID learn more from evaluations.

References

Aarons GA (2012) Mixed methods for implementation research: application to evidence-based practice implementation and staff turnover in community-based organizations providing child welfare services. *Child maltreatment p* 17(1): 67–79.

Abadie A (2010) A. Diamond, and J. *Hainmueller Synthetic control methods for comparative case studies Estimating the effect of Californias tobacco control program Journal of the American statistical Association p* 105(490): 493–505.

Befani B and Stedman-Bryce G (2017) Process Tracing and Bayesian Updating for Impact Evaluation. *Evaluation* 23(1): 42–60.

Befani B, Stefano D'Errico, Booker F, et al. (2016) *Clearing the Fog: New Tools for Improving the Credibility of Impact Claims, in Paper.*

Bernal JL (2017) S. Cummins, and A. *Gasparrini Interrupted time series regression for the evaluation of public health interventions a tutorial International journal of epidemiology p* 46(1): 348–355.

Bessette J (1980) Deliberative democracy: The majority principle in republican government. *How democratic is the constitution p* 102: 109–111.

Blanchet K (2017) Evidence on public health interventions in humanitarian crises. *The Lancet*.

Bonell C (2012) Realist randomised controlled trials: a new approach to evaluating complex public health interventions. *Social science medicine p* 75(12): 2299–2306.

Bonell C and J (2014) Dark logic': theorising the harmful consequences of public health interventions. *Community Health p jech* 2014: 2014–204671.

Bonell CP (2008) Alternatives to randomisation in the evaluation of public-health interventions: design challenges and solutions. *Journal of Epidemiology Community Health p jech 082602* 2009.

Breuer E (2016) Using theory of change to design and evaluate public health interventions: a systematic review. *Implementation Science p 63* 11(1).

Bryman A (2006) Integrating quantitative and qualitative research: how is it done? Qualitative research, . : p. . *Qualitative Research* 6(1): 97–113.

Burchett HE (2013) The applicability and transferability of public health research from one setting to another: a survey of maternal health researchers. *Global health promotion p* 20(1): 16–24.

Cartwright N (2011) A philosopher's view of the long road from RCTs to effectiveness. *The Lancet p* 377(9775): 1400–1401.

Carvalho S and White H (1997) Combining the quantitative and qualitative approaches to poverty measurement and analysis: the practice and the potential. *World Bank Publications* 23.

Chan A-W, Tetzlaff JM, Altman DG, et al. (2013) SPIRIT 2013 Statement: Defining Standard Protocol Items for Clinical Trials. *Annals of Internal Medicine* 158(3): 200–207.

Chianca T (2008) The OECD/DAC criteria for international development evaluations: An assessment and ideas for improvement. *Journal of Multidisciplinary Evaluation p* 5(9): 41–51.

Craig P (2008) Developing and evaluating complex interventions: the new Medical Research Council guidance. *Bmj p a1655* 337.

Creswell JW (2011) *Best practices for mixed methods research in the health sciences. National*

Institutes of Health, Bethesda (Maryland).

Daniel RM (2011) B.L. De Stavola, and S. *N Cousens gformula Estimating causal effects in the presence of timevarying confounding or mediation using the gcomputation formula Stata Journal p 479* 11(4).

Davey C, Boulay M and Hargreaves JR (2014) Strengthening Non-Randomized Studies of Health Communication Strategies for HIV Prevention. *R Hargreaves Strengthening nonrandomized studies of health communication strategies for HIV prevention of Acquired Immune Deficiency Syndromes p* 66: S271–S277.

Davies R (2004) Scale, Complexity and the Representation of Theories of Change. *Evaluation* 10(1): 101–121.

DFID (2014) *Can agriculture interventions promote nutrition? DFID*.

Gertler PJ, Sebastian M, Premand P, et al. (2011) *Impact Evaluation in Practice*. The World Bank.

Glasgow RE (1999) T.M. Vogt, and S. *M Boles Evaluating the public health impact of health promotion interventions the REAIM framework American journal of public health p* 89(9): 1322–1327.

Hahn J (2001) P. Todd, and W. *Van der Klaauw Identification and estimation of treatment effects with a regressiondiscontinuity design Econometrica p* 69(1): 201–209.

Hey SP (2015) What Theories Are Tested in Clinical Trials? Philosophy of Science, . : p. . *Philosophy of Science* 82(5): 1318–1329.

Humphreys M (2015) Mixing Methods: A Bayesian Approach. *The American Political Science Review* 109(4): 653–673.

Manandhar DS (2004) Effect of a participatory intervention with women's groups on birth outcomes in Nepal: cluster-randomised controlled trial. *The Lancet p* 364(9438): 970–979.

Marchal B (2013) Realist RCTs of complex interventions-an oxymoron. *Social Science Medicine p* 94: 124–128.

McDermott R (2011) *Internal and external validity. Cambridge handbook of experimental political science p*.

Mohr LB (1995) *Impact Analysis for Program Evaluation*. Sage.

Moore Graham F, Suzanne A, Mary B, et al. (2015) Process evaluation of complex interventions: Medical Research Council guidance. *BMJ* 350: 1258.

Oakley A (2006) Health services research: process evaluation in randomised controlled trials of complex interventions. *BMJ British Medical Journal p 413* 332(7538).

Palinkas LA (2011) Mixed-methods designs in mental health services research: A review. *Psychiatric Services p* 62(3): 255–263.

Pawson R and Tilley N (1997) *N. Tilley, Realistic evaluation.* Sage.

Pearl J (2011) E. Bareinboim. Transportability of causal and statistical relations: A formal approach. *in Data Mining Workshops ICDMW IEEE 11th International Conference on IEEE*.

Petticrew M (2013) Synthesizing evidence on complex interventions: how meta-analytical, qualitative, and mixed-method approaches can contribute. *Journal of clinical epidemiology p* 66(11): 1230–1243.

Popper KR (2002) *The poverty of historicism. Psychology Press*.

Prost A (2017) *Facilitated participatory and action groups to improve maternal and newborn health at scale in Jharkhand, India*.

Ragin CC (1999) Using qualitative comparative analysis to study causal complexity. *Health Serv Res* 34(5 pt 2): 1225–1239.

Richiardi L (2013) R. Bellocco, and D. *Zugna Mediation analysis in epidemiology methods interpretation and bias International journal of epidemiology p* 42(5): 1511–1519.

Roche C (2015) What has cancer taught me about the links between medicine and development? *From Poverty to Power, Oxfam*. Available from: https://oxfamblogs.org/fp2p/what-has-cancer-taught-me-about-the-links-between-medicine-and-development-guest-post-by-chris-roche/ (accessed 1 May 2017).

Rosen L (2006) In defense of the randomized controlled trial for health promotion research. *American Journal of Public Health p* 96(7): 1181–1186.

Rosenzweig M and Udry C (2016) *C. Udry, External validity in a stochastic world.*

Rossi PH, Lipsey MW and Freeman HE (2003) *Evaluation: A Systematic Approach*. *Sage publications*.

Shadish WR, Cook TD and Campbell DT (2002) *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. *Experimental and Quasi-experimental Designs for Generalized Causal Inference*, Michigan: Houghton Mifflin.

Shaffer JP (1995) Multiple hypothesis testing. *Annual review of psychology p* 46(1): 561–584.

Sharma KR (1999) Farm commercialization and nutritional status of children: the case of the vegetables, fruits, and cash crops programme in western Nepal. *Food and Nutrition Bulletin p* 20(4): 445–453.

Silva MJ (2014) De , Theory of Change: a theory-driven approach to enhance the Medical Research Council’s framework for complex interventions. *Trials p 267* 15(1).

VanderWeele TJ (2010) S. Vansteelandt, Odds ratios for mediation analysis for a dichotomous outcome. *American journal of epidemiology p* 172(12): 1339–1348.

Wang R (2007) Statistics in medicine—reporting of subgroup analyses in clinical trials. *New England Journal of Medicine p* 357(21): 2189–2194.

Weiss CH (1997) Theory-based evaluation: Past, present, and future. *New directions for evaluation p* 1997(76): 41–55.

White H (2002) Combining quantitative and qualitative approaches in poverty analysis. *World Development* 30(3): 511–522.

White H and Phillips D (2012) *Addressing attribution of cause and effect in small n impact evaluations: towards an integrated framework*. *International Initiative for Impact Evaluation.Working Paper 15*.

Contact us

Centre of Excellence for Development Impact and Learning
London International Development Centre
36 Gordon Square
WC1H 0PD
www.cedilprogramme.org
@CEDIL2017