Towards a multi-dimensional model of impact evaluation quality: assessing development impact evaluation methods with respect to context

Matthew Tom Juden

Thesis submitted for the degree of PhD

2021

Department of Development Studies

SOAS, University of London

ABSTRACT

In this thesis, I give an account of the quality of (quasi-)experimental impact evaluation methods for development interventions that goes beyond internal validity considerations to also incorporate the transferability of findings to new contexts. To investigate how well different (quasi-)experimental impact evaluation methods facilitate transferability, I adapt tools from realist synthesis to extract and synthesise the programme theories that underpin a set of evaluations of the same intervention-outcome pair. From this synthesis of programme theory, I derive the markers of intervention causation in context (MICCs) that a (quasi-)experimental impact evaluation of an intervention of that type would have to report to facilitate the transferability of findings. I systematically build a complete set of (quasi-)experimental impact evaluations for two intervention-outcome pairings, and apply my method to these two cases. This generates case-specific insights such as identifying evidence gaps where minimal further data generation offers large gains in understanding. Further, the analysis generates cross-case insights such as the tendency to better report causally significant features of intervention implementation than causally significant features of context. Most importantly, the analysis suggests there is no association between method choice and the facilitation of transferability, in theory or in practice. I argue that we can nonetheless improve on 'there is no gold standard' by showing how generating a middle-range theory of intervention causation capable of underpinning the list of MICCs for a type of intervention provides a guide to evaluation method choice and to transferring results between contexts.

In parallel, to render my main results more useful to the relevant experts and therefore more likely to influence practice, I conduct semi-structured interviews with development intervention evaluation experts. I identify a broad discursive trend in favour of theory-based evaluation and a nascent interest in the use of middle-range theories to underpin transferability. I therefore frame my main results in these terms.

ACKNOWLEDGEMENTS

I would like to thank the Economic and Social Research Council for funding this work. Professor Peter Mollinga was invaluable in encouraging me to pursue this project and in supervising me. I would also like to thank Professor Chris Bonell and Dr Barbara Befani for a helpful viva conversation and for corrections which, although ‘minor,’ substantially improved this thesis in key areas.

Thanks are very much due to my partner Camille, without whom I would never have completed the PhD or many other projects besides. I would also like to thank Zéphir for arriving at an inconvenient but nonetheless perfect moment, and my parents for saving the day.

CONTENTS

1 Introduction

In this thesis I begin from the observation that the practice of impact evaluation of development interventions has a problem. The problem begins with ‘gold standard’ thinking about randomised controlled trials (RCTs), which places them at the top of a hierarchy of impact evaluation methods, and places large-*N* impact evaluation at the top of a hierarchy of types of methods that devalues other ways of learning about development. This problem is compounded by a widespread acceptance that there is a ‘problem of external validity’ that undermines the special value of (quasi-)experimental impact evaluation in general and RCTs in particular, creating a nascent methodological crisis. Philosophers of science have been quick to highlight the problem of external validity and to argue that ‘there is no gold standard.’ However, these arguments have done little to change practice. This is perhaps because they have been perceived to amount to ‘anything goes,’ leaving no clarity around how else the quality of methods should be assessed, how methods should be chosen, and how results issued from different methods should be interpreted.

As a researcher in development I am driven, like many, by the hope that our research can contribute to policy change which improves the conditions of the poor and marginalised. While I am under no illusions that development policy (public and private) is entirely ‘evidence-based,’ I begin from the assumption that it is to some extent evidence-informed, and that it is this small influence on policy that makes research valuable. To the extent that this assumption holds true, a problem in the selection, design and interpretation of research methods is a social problem worth working to improve.

In this thesis I do not presume to develop an account of evidence quality or method choice in general; that is clearly out of scope and may be incoherent. Rather, I limit my scope to impact evaluations assessing the effectiveness of development interventions. All the above is discussed in much more detail in Chapter Two of this thesis, which reviews the literature to motivate my primary research question, which is:

> *Can we give a useful, systematic account of the relative merits of evidence generated using different (quasi-)experimental development impact evaluation methods that goes beyond internal validity to also consider the extent to which methods facilitate the transfer of results to other contexts? If so, how?*

The astute reader will note that the ‘problem of external validity’ was referred to above, but that the primary research question is framed in terms of the ‘transfer of results to other contexts.’ I began work on this thesis using ‘external validity’ as a lens for my analysis, but careful thinking about what ‘external validity’ meant to the many different authors in the literature revealed the term to be confused. In Chapter Three, Section 3.1, I set out this thinking and argue that ‘external validity’ has at least three incompatible meanings in the way it is used in the literature. I distinguish these three different meanings by labelling them ‘transferability’ (the extent to which treatment effects will hold in some other context), ‘generalisability’ (the extent to which treatment effects from some sample context will hold in the population context) and ‘observed heterogeneity’ (the extent to which given treatment effects are predictive of the treatment effect across all observed contexts). The ‘problem of external validity’ as identified by many authors in the literature is more precisely a problem of transferability. I therefore state my primary research question using this more precise term.

To answer the primary research question, I must compare the merits of (quasi-)experimental impact evaluation methods regarding the extent to which they facilitate the transfer of their results to other contexts. Chapter Three sets out my theoretical approach, arguing that I am forced to develop a novel method. It is the development and use of this method that forms the heart of this thesis. I adapt programme theory mapping from realist synthesis to identify the markers of intervention causation in context (MICCs) that must be generated and reported by evaluations of a type of intervention in order to facilitate arguments for the transferability of their findings. I can then use this method for two case studies of sets of evaluations of the same intervention-outcome pair to investigate any differences in the reporting of MICCs between evaluations using different (quasi-)experimental impact evaluation methods. The case studies examined are all the (quasi-)experimental impact evaluations systematically identified of 1) conditional cash transfers for school enrolment and 2) deworming for child weight. I chose the case studies on the basis that they were the best-studied, with the most (quasi-)experimental impact evaluations available. In addition, these case studies are helpfully contrasting and representative allowing more to be learned about development evaluations in general from analysis and comparison of the two cases. Sabet and Brown (2018) estimate that roughly 50% of evaluations of development interventions are published by social scientists and economists and the remainder are published by public health researchers and epidemiologists. The two cases studied in this thesis straddle this divide, with one set of evaluations being published overwhelmingly in health journals, and the other set overwhelmingly in economics and social science journals or as working papers. Furthermore, one intervention is a public health intervention and the other straddles social protection and education. Approximately 65% of all interventions evaluated in the (quasi-)experimental development impact evaluation literature are conducted in one of these three areas (Sabet and Brown, 2018).

Chapters Five and Six describe for each case in turn how the sample of evaluations was identified and how the programme theory underpinning evaluations in each set was extracted and synthesised. The process of moving from programme theory to generating the lists of MICCs is also described for each case. These chapters are a sort of narrative synthesis of the programme theories active in each literature that are informative in their own right. Analysing the reporting of MICCs by evaluations in both sets generates further useful insights for those literatures that are discussed in Chapter Seven. For example, non-financial barriers to education such as erroneously low estimates of the returns to education or failures of rationality are widely theorised to be a barrier to enrolment that might condition the effectiveness of a conditional cash transfer program. However, the MICCs associated with these contextual features are reported by very few evaluations, representing a gap in the literature and in our understanding of contextual determinants of the effectiveness of CCTs. Similarly, entry of soil-transmitted helminths (STHs) into the body via a bare foot is widely acknowledged to be a key vector for STH transmission. Despite this, only one study in the set of evaluations in Case Two reports whether children wear shoes in the community targeted for the intervention. This means that the moderating effect of child footwear may or may not be responsible for much of the variation in effectiveness of deworming interventions. The fact that almost no evaluation reports this key moderating variable for the effectiveness of the intervention suggests that the evidence base could be enriched substantially by revisiting existing evaluations and attempting to generate this data for the communities targeted by the interventions studied. It also means that experimental testing of the moderating effect of footwear on the effectiveness of deworming should similarly enrich the available evidence base. That such a glaring omission has gone unremarked and uncorrected in the literature is an inditement of (quasi-)experimental impact evaluation practice and is further evidence that a focus on the unbiasedness of treatment effect estimates to the exclusion of other quality considerations has impoverished the evidence base in the study of development. The preceding insights, and the rest of the argument of Chapter Seven, demonstrate that my novel method, the generation and analysis of MICCs, is a success, generating informative insights for evaluation practitioners and evidence synthesisers working in the literatures studied.

With the method determined to be generating useful, informative insights, rather than just noise, I turn in Chapter Nine to assessing what has been learned about the extent to which different (quasi-)experimental impact evaluation methods facilitate transferability of their findings. First, Chapter Eight reports the results of a parallel strand of research. This research was conducted in response to the demand in my primary research question that the account developed be not only systematic but *useful*. Having noted in my literature review that valid philosophical arguments are not sufficient to change practice, I set out to determine how the account developed could be framed so as to be most useful to experts on the (quasi-)experimental impact evaluation of development interventions. I conducted semi-structured interviews with a sample of these experts using a critical realist epistemic communities approach to attempt to achieve three objectives: 1) identify epistemic communities that claim authority over judgements of the quality of development intervention evaluation evidence; 2) identify their ‘shared notions of validity’ concerning what counts as a ‘high quality’ (quasi-)experimental impact evaluation; and 3) identify the features of these accounts that are valued by members of the community as well as any unresolved puzzles or nascent crises that put pressure on them. An iterative movement between the data generated by these interviews and the relevant literature reveals a picture of two closely linked epistemic communities: the *randomistas* and the sceptics.[1] While being divided over whether RCTs should have a special place atop a hierarchy of impact evaluation methods, these communities share a frustration with several nascent crises in the development impact evaluation literature. Members of both communities endorse an emergent new hegemony in the evaluation of development interventions, which is referred to in the literature as ‘theory-based evaluation’. This framework remains extremely vague in the literature; although it is widely endorsed it is not at all clear what is required of evaluators in practice. Similarly, ‘middle-range theory’ is increasingly seen to be the key to transferability by practitioners at the cutting edge of (quasi-)experimental impact evaluation, but it is not yet clear how such theories can be generated and how they can be used. This situation presents an enormous opportunity for the answer to the primary research question generated in this thesis to be useful, and Chapter Nine therefore presents that answer in terms that respond to the demand to put flesh on the bones of theory-based (quasi-)experimental impact evaluation and to clarify the utility of middle-range theory.

In the first sections of Chapter Nine, several lessons are drawn for the development impact evaluation literature based on the analysis of evaluations in both sets. Chief amongst these lessons is that the MICCs are not sufficiently investigated and reported by evaluations from either set. This is especially true for that subset of MICCs that deal with features of the pre-existing study context (like prevalence of shoe-wearing for de-worming evaluations). Features of intervention design (like the nature of the conditionality for a conditional cash transfer

[1] I adopt the term ‘*randomista*’ despite the fact that some, such as Webber and Prouse (2018, p.4) consider it to be a ‘gendered, derogatory term intended to flippantly dismiss experimental economists and their success, particularly Esther Duflo.’ I employ the term as it is in very widespread use, including often as a self-description. However, it is important to note the criticism and to be clear that I do not intend any derogatory connotation.

programme) are reported more often, though still not by all evaluations in the sets studied. In the latter sections of the chapter, I draw together all the strands of this research project and develop my answer to the primary research question. I argue that there is no relationship, contingent or necessary, between method choice and the transferability of results. Analysis of the evaluations in both sets shows no relationship between method choice and the ability of interventions to generate and report MICCs, nor is any such relationship suggested by theorising about how each method might ideally be used. This may appear inconsistent with widespread belief in the trade-off between internal and external validity. However, my analysis reveals this apparent inconsistency to be an artefact of the confused way in which 'external validity' is used as an umbrella term that encompasses at least three quite different meanings. These were referred to earlier in this introduction and are discussed more in Chapters Three and Nine.

It might seem, then, that this leaves us 'back where we started,' so to speak, in that (quasi-)experimental impact evaluation methods have been found not to generate more or less transferable insights than each other either in theory or in practice. If some methods are more able to generate more reliably internally valid estimates than others, then perhaps this forces us to re-embrace a uni-dimensional account of evidence quality based on internal validity. However, this is not the case. In the latter sections of Chapter Nine I argue that generating a middle-range programme theory of the sort of intervention concerned can provide the basis for a systematic approach to method choice and to the transfer of results between contexts. I also show that there are additional benefits for internal validity of generating the list of MICCs for an intervention-outcome pair being studied. Finally, I argue that, although this thesis makes use of a realist ontology and epistemic strategy, embracing the philosophy and tools of Realist Evaluation is not necessary to generate middle-range programme theories and lists of MICCs. I discuss work published by myself and co-authors including Nancy Cartwright that develops an alternative approach framed in terms of what John Stuart Mill called 'tendency principles.' This renders the principles for method choice and the interpretation of (quasi-)experimental impact evaluation results from other contexts developed in Chapter Nine accessible for those who are reluctant to embrace capital R Realism, as the analysis of Chapter Eight suggests some might be.

This thesis is presented in two parts. Part One describes my motivation and approach, including three chapters that comprise a literature review, theoretical approach and discussion of methodology. Part Two describes lessons learned. It comprises five chapters; one for a description of the application of the method to each case, a chapter describing the insights generated for both cases, a chapter detailing findings from the expert interviews, and a final chapter answering the primary research question. There then follows a conclusion, references and appendices.

Part one: motivation and approach

2 Literature survey

This chapter begins with a presentation of the primary research question to which this thesis responds. The rest of the chapter is dedicated to motivating this research question. It begins by describing the rise of ‘evidence-based policy’ rhetoric in development policy and explains that this discourse encourages a view of randomised controlled trials as the ‘gold standard’ in impact evaluation. In Section Two, the strengths of experimental and quasi-experimental methods are reviewed. Section Three presents the ‘problem of external validity’ that undermines gold standard thinking. Section Four argues that decrying gold standard thinking is not sufficient without providing an alternative way of satisfying the intuition that some methods are better than others. The primary research question is thus motivated.

Primary research question:

> *Can we give a useful, systematic account of the relative merits of evidence generated using different (quasi-)experimental development impact evaluation methods that goes beyond internal validity to also consider the extent to which methods facilitate the transfer of results to other contexts? If so, how?*

2.1 THE RISE OF ‘EVIDENCE-BASED POLICY’ RHETORIC IN DEVELOPMENT POLICY

Researchers have been explicitly attempting to influence public policy at least since Guerry tried to show that education did not reduce crime in 1833 (Weiss, 2009, p.ix, citing Cullen, 1975, p.139). More recently, policy-makers have begun to claim that they seek to act on the basis of evidence. Perhaps the most famous example of this rhetoric is the ‘modernising government’ agenda of the Labour government in power in the United Kingdom from 1997 to 2010, which sought to ‘improve [government’s] use of evidence and research’, emphasising evidence of ‘what works’ (Cabinet Office, 1999, sec.2). This sort of rhetoric is now very widely employed, with repeated commitments at the highest level of policymaking across the world to make public policy more ‘evidence-based’.[2] In the last four years, the election of Donald Trump in the United States, and the rise to power of more populist governments in Brazil, Hungary, Indonesia, and to a limited extent the United Kingdom has seen some political players in public discourse turn against ‘evidence’. Famously, in 2016 British member of parliament and Secretary of State for Justice Michael Gove declared that the public had ‘had enough of experts with organisations with acronyms saying that they know what is best and getting it consistently wrong.’ Nevertheless, the effect of this change in discourse on actual policymaking should not be overstated. The drive to produce ‘evidence-based policy’ continues to play an important role in the thinking of policymakers. Institutional arrangements designed to increase the role of evidence and set up before 2016 persist even in the US and UK.

Let us consider development policymaking in particular: talk about ‘the evidence’ has come to be central to development policy discussions. This is especially true of ‘development policy’ as it relates to the policies of advanced economies when deploying their overseas development assistance budgets. It is also true to a lesser extent, to some extent due to the influence of the Bretton Woods institutions, to public policy made in developing countries (Carden, 2009, pp.4–7; Weiss, 2009, pp.x–xi). For example, the UK Foreign, Commonwealth and Development Office (FCDO), claims to deliver its programmes in accordance with 10 principles, one of which is that programmes should always be ‘evidence-based’ (FCDO, 2020, p.9). Further, since 2011, all of the UK’s use of official development assistance (ODA or ‘aid’) has been scrutinised by the Independent Commission for Aid Impact (ICAI), who report their assessments directly to the International Development Select Committee. ICAI’s work is guided by five ‘core values’, one of which is that they are ‘committed to the rigorous use of evidence and analysis in our reviews’ (ICAI, 2020, p.1).

The prevalence of the rhetoric of evidence-based policy is not sufficient to imply that policy decisions are entirely or even partially determined by appraisal of the available evidence. The

[2] See, for example, Deans and Ademokun (2013) for examples from Australia and Tanzania, Gluckman (2013) for an example from New Zealand, Broadbent (2012) for examples from Ghana, Uganda, Zambia and Sierra Leone.

overwhelming consensus in the literature is that the interaction between factors influencing policy decisions creates a complex process in which research findings have only limited power to influence policy (Brownson et al., 2006; Jones and Villar, 2008; Keck and Sikkink, 1998). Even when policy-makers seek to make use of research, they report barriers to this attempt, especially in developing country contexts (Hyder et al., 2011). Furthermore, the power of evidence to influence policy is extremely difficult to model. As Maxwell and Stone (2006, p.1) put it: '[a] linear model, in which careful research leads inexorably to better policy, is widely derided'.

Nevertheless, research evidence can have an impact on policy. The work of Haas (1992) and many others catalogues examples of research promoted by expert communities influencing government learning and changing government policy. In development specifically, Carden (2009) for the IDRC presents a set of case studies of research outputs that have influenced public policy in developing countries. Court and Young (2006) explore examples of research influencing policy at the World Bank, the IMF and other multilateral donor organisations, as well as at DFID.

The rise of evidence-based policy rhetoric has certainly had an impact on development research practice. This process is also complex, and precise channels of causation are hard to identify. However, there is a general agreement in the literature that the primary importance of results and value-for-money in a policy development process that should be based on the 'best evidence' has created the conditions for the rise of research quality assessment frameworks issued from the natural sciences and assimilated via public health research.[3] These research quality assessments are based on the absolute primacy of 'internal validity'. That is, the extent to which the research design employed minimises problems of confounding on unobservable characteristics of treatment units to facilitate confident causal attribution of changes in outcomes to the action of the intervention being studied.

[3] See, for example, Eyben *et al.* (2015) for a collection of persuasive pieces that explore different aspects of this process. See also Langer and Stewart (2014) for a review of the changes in institutional priorities that led to the rise of medical-style systematic reviewing in development.

Frameworks for the assessment of any research evidence in terms of internal validity have been explicitly stated. For example, the Maryland Scientific Methods Scale gives research evidence issued from different methods a score of one to five, dependent entirely on the extent to which the method employed supports a judgement of internal validity of the results (Farrington et al., 2002). These scores can be downgraded by one point in the presence of serious problems with attrition, for example, but it is impossible for a study that does not employ 'at least' a quasi-experimental matching design to score more than three points. Randomised controlled trials (RCTs) are the only designs that can score 5 points. Though based on much older ideas originally articulated by Cook and Campbell (1979), the Maryland Scientific Methods Scale was developed in 2002 and remains in use today, for example by the What Works Centre for Local Economic Growth in the UK (Puttick, 2018).

Across the community of experts that claim authority over what counts as the 'best' evidence of relevance to development policy, there is a tendency to employ a hierarchy of methods that values experimental methods above quasi-experimental methods, those more than observational methods, and systematic reviews of experimental and quasi-experimental studies most of all (Mallett et al., 2012; Eyben et al., 2015; Langer and Stewart, 2014; Cameron, Mishra and Brown, 2016). Within these frameworks, the randomised controlled trial is considered to be the 'gold standard' best possible study design, and other methods are judged solely on their ability to approximate the form of the RCT (Deaton, 2010).

It is more common for frameworks based solely on internal validity to remain somewhat implicit than for them to be actively stated. For example, a 2020 review paper from academics at the Abdul Latif Jameel Poverty Action Lab (J-PAL) is typical of the 'what works' literature in that it sets out to answer the question 'What works to enhance women's agency?' (Chang et al., 2020, p.1). However, the review methodology only allows the authors to examine 'quality' studies and this is explained as follows: 'In addition to RCTs, we included studies that used quasi-experimental designs with well-tested assumptions, including difference-in-differences (DID), instrumental variables, and regression discontinuity (RD)' (*ibid*, p.11). By presenting this review as a summary of ‘what works,’ the authors implicitly accord no epistemic value to matching or purely observational studies. It is clear in their presentation of the inclusion criteria that the authors hold RCT evidence to be the ‘gold standard’ and measure other methods by the extent of their deviation from that standard.

It is the frameworks increasingly being employed for judging evidence quality that are the object of this research, rather than the research to policy process as a whole. It is my hope that by introducing a little realist anti-discipline to the assessment of prevailing, implicitly positivist frameworks, the practice of development experts can be changed so as to represent the quality of the available evidence with more nuance. More detail on this theoretical framework will be presented in Chapter Three.

2.2 THE APPEAL OF EXPERIMENTAL AND QUASI-EXPERIMENTAL METHODS

In part, the rise of experimental and quasi-experimental methods in development is attributable to shifts in the thinking of actors on the supply side of research, as well as to the situation on the demand side described in the previous section. Picciotto (2012) persuasively describes a crisis of development economics around the turn of the century, resultant from ambiguous results of unclear evaluations of the effectiveness of aid spending. In particular, the local average treatment effect (LATE) problem with instrumental variables approaches was understood to undermine many of the World Bank’s claims for programme effectiveness (Ogden, 2016, pp.55–56). This crisis created the conditions in which Duflo and Kremer (2005, p.228) were able to declare, during the 2003 World Bank Conference on the evaluation of development effectiveness, that ‘[j]ust as randomized evaluations revolutionized medicine in the 20th century, they have the potential to revolutionize social policy during the 21st’.

The call-to-arms issued by Duflo, Kremer, Banerjee and others (later to be dubbed the *randomistas*) was persuasive in part because the internal validity of research evidence is extremely important. It would be implausible to suggest that the rise of the systematic review, randomised controlled trial, and quasi-experimental methods has nothing to do with the power of these methods in addressing the research questions to which they are suited. In particular, like instrumental variables approaches, these methods offer a way of overcoming problems of endogeneity, in which independent variables (i.e. participation in the intervention) are suspected to be a function of dependent variables (i.e. the outcome of interest) (White, 2011). However, unlike instrumental variables approaches, they generate estimates of treatment effects for the whole trial population, rather than a LATE for the subset of participants for whom the instrument determined their treatment.

The unique strength of RCTs, as Bonell *et al.* (2012, p.2300 emphasis added) put it, is that they 'generate minimally biased estimates of intervention effects by ensuring that *intervention and control groups are not systematically different* from each other in terms of measured and/or unmeasured characteristics'. This is not quite literally true. As Deaton and Cartwright (2018b) remind us, this is only true *in expectation,* and is not likely to be true *in fact* for any single trial. Nevertheless, as the sample size of the trial increases, deviation from this assumption of balance becomes less likely. Further, a variety of matching or stratification techniques can be used pre-randomisation to guarantee a minimal level of balance on observed covariates, or covariates can be used to constrain the process of randomisation in various ways (Ivers et al., 2012). Quasi-experimental approaches also generate minimally biased estimates of treatment effects, though on a larger set of assumptions. Randomness is used to overcome problems of endogeneity by such approaches, too, albeit a randomness that is identified by a researcher rather than created by them. For example these sorts of randomness include living just to one side of a line on a map or the other, or achieving a test score just slightly higher or lower than the a cut-off point (Hahn, Todd and Van der Klaauw, 2001).

Employing a design that minimises systematic differences between beneficiaries (of various forms of an intervention) and non-beneficiaries of an intervention is a powerful way of answering questions of the form 'what would have happened if beneficiaries had not been exposed to (a particular form of) the intervention?' The power of experimental and quasi-experimental designs to answer these questions comes from making such questions as close as possible to equivalent to questions about what happened to the relevant group of people in the study. Specifically, experimental and quasi-experimental approaches minimise the demands of the set of assumptions required to make those two sets of questions logically equivalent (Cartwright, 2007).

Criticisms of experimental and quasi-experimental methods from within development economics have occasionally focussed on undermining their supposed superiority with regards to internal validity. Heckman (1991), Worral (2007), and Deaton (2010), for example, outline a host of concerns. Chief among these are the facts that failure to 'blind' subjects and researchers undermines the precision of estimates; that experimental and quasi-experimental designs are informative about mean treatment effects, but do not identify other features of the distribution; and that statistical inference in experiments and quasi-experiments is much more complex than it seems. However, Banerjee and Duflo (2008), are persuasive in their argument that experimental and quasi-experimental methods address these problems at least as well as comparable methods. Cartwright and Deaton (Deaton and Cartwright, 2018b) raise several more problems with the way in which decisions to randomise are made and with the way in which RCT results are interpreted. However, nothing that they say is intended to argue that the RCT is not a powerful tool. As they say (*ibid*, p.3) 'we are not against RCTs, only magical thinking about them.'

Given their unique strength in isolating the effects of interventions from other potential causes, experimental and quasi-experimental methods are indisputably powerful tools for uncovering the effects of development interventions. Nonetheless, there is an enormous problem with the uni-dimensional measures of evidence quality, imported from medicine, that have accompanied the growth in popularity of experimental and quasi-experimental methods. This is the so-called 'problem of external validity'.

2.3 THE PROBLEM OF EXTERNAL VALIDITY

The great strength of experimental and, to a lesser extent, quasi-experimental methods, is that the validity of their findings for the study population do not rely on a fallible model of how intervention causation works within that study population. To borrow Cartwright's (2007) terminology: in the case of an RCT, we do not need to know *how* intervention T causes outcomes O in population φ in order to deduce *that* it did; so long as we have conducted a good RCT and can rely on the assumption that there are no systematic differences between treatment and control groups within φ. This is what Deaton (2010, p.28) refers to as 'the magic that is wrought by the randomization'.

In quasi-experimental methods, strategies other than randomised allocation to the intervention are used to attempt to construct control groups of non-beneficiaries that are as similar as possible to beneficiaries. These methods differ in the amount of theorising about intervention causation that is involved. For example, in the case of regression discontinuity designs, subjects are assigned to the intervention or not based on their score on some continuous, normally distributed, variable, with the probability of assignment jumping discontinuously to 1 from 0 at some cut-off point (Hahn, Todd and Van der Klaauw, 2001). The most obvious example of such a situation is an intervention in which assignment to the intervention depends on test scores, with subjects over a certain score receiving the intervention, and others not. In this case, the groups of subjects that were *just over* the cut-off score are compared with subjects that were *just under*. Almost no fallible theorising about intervention causation is required to deduce that the overwhelming majority of the difference between the two groups in terms of outcomes is attributable to the intervention. This is because we can be confident that there are extremely small systematic differences between the two groups.

By contrast, matching methods are observational methods that do require a fairly high level of theorising about intervention causation in order to construct a control group and thereby test the counterfactual. In such studies, one of several possible methods is used to match subjects who did receive the intervention with subjects that did not, but who are otherwise extremely similar in terms of the characteristics that are theorised to be relevant to intervention causation (Stuart, 2010). For proponents of the frameworks for judging evidence quality discussed above, the greater the level of theorising about intervention causation that is involved in the construction of a quasi-experiment, the lower the level of internal validity its results are judged to have.

The aversion to theorising about intervention causation discussed above becomes a problem for proponents of uni-dimensional accounts of evidence quality once the need to apply results from one context to another is taken into account. As discussed above, experimental and quasi-experimental methods can allow researchers to deduce that intervention T causes outcomes O in population φ with little or no examination of the causal structure of φ. However, applying these results to some target population θ requires an argument that the causal structure of θ is relevantly similar to that of φ. A study that has left the causal structure of φ unexamined cannot provide premises for such an argument (Cartwright, 2007, 2008). This is what is meant by the problem of 'external validity'. The great strength of methods that do not rely on theorising about causes becomes their great weakness. Basu (2013) and Deaton (2010) point out that this problem also applies to population φ at a different time and to any non-random subset of φ. The fact that such populations are not 'external' to the original study population has led scholars such as Astbury and Leeuw (2010) to refer to a more general 'black box problem' in place of a problem of 'external validity'.

Emerging empirical evidence suggests that the problem of external validity cannot be waved aside based on an assertion that, in fact, the populations of interest to development scholars are proving to be relevantly homogeneous. Vivalt (2020) constructs a dataset of 'impact evaluations' (experimental or quasi-experimental studies of the impact of a development intervention) and examines the heterogeneity of results. There are many ways of interpreting the presented heterogeneity. However, one pertinent finding is that the median probability that looking at the sign of a (quasi-)experimental impact evaluation would correctly predict the sign of another (quasi-)experimental impact evaluation of the same pairing of intervention and outcome was 61% (Vivalt, 2020, p.23). While Vivalt (2016, p.26) is cautious in her interpretation of this and other measures of heterogeneity presented, she concludes that 'it is safe to say that these (quasi-)experimental impact evaluations exhibit more heterogeneity than is typical in other fields, such as medicine.' Her research also provides an opportunity to investigate some of the common correlates of heterogeneity across studies, providing unsurprising but nonetheless valuable evidence that smaller studies and studies of academic or

NGO-implemented interventions tend to have larger effect sizes than do larger studies and studies of interventions implemented by governments. These findings provide empirical support for the Ravallion's (2009, p.33) warning that it was significant and potentially detrimental to the quality of the evidence base that 'randomization is only feasible for a nonrandom subset of the interventions and settings relevant to development'.

It might be hoped that combining theory and empirical results would allow for much-improved estimates of the likely generalisability of results from an experiment in one setting to another. However, in their examination of the theoretical models and empirical evidence on returns to schooling, Pritchett and Sandefur (2013, p.29) have demonstrated 'that even for an economic model that has been studied *ad nauseum*, we are not currently in a strong position to combine theory and empirics to make externally valid claims about parameters' magnitude. If you want to know the return to schooling in country X, there is no reliable substitute for data from country X.' Pritchett and Sandefur extend this analysis in a 2015 paper, which analyses two primary outcomes from a microcredit intervention implemented fairly consistently in six different contexts (Pritchett and Sandefur, 2015). They show that for these evaluations, selection-biased observational estimates from the same context are less biased than experimental evidence from another setting. This remains the case even when aggregating experimentally estimated treatment effects from up to three experiments for one outcome of interest and up to five experiments for the other.

The problem of external validity, then, coupled with the observed heterogeneity of results of (quasi-)experimental impact evaluations of development interventions, provides a fatal challenge to uni-dimensional assessments of evidence quality based on internal validity alone. Take, for example, a hypothetical government policymaker considering policy options for a given context. It is very far from clear that *observational* data from a large-scale government-implemented programme in a similar context is less useful for this policy maker than *experimental* data from a small-scale NGO-implemented programme in a less similar context

(Pritchett and Sandefur, 2015; Vivalt, 2020). *Contra* the *randomistas*, which evidence is 'best' in such a situation is at least open to interpretation.

2.4 THE POVERTY OF 'THERE IS NO GOLD STANDARD'

So, elevating the ability to derive maximally internally valid, minimally biased estimates of mean treatment effects above all other criteria for assessing the quality of evidence is misguided, and philosophers of science have been assiduous in pointing this out. When it comes to assessing evidence quality for public policy questions, philosophers of social science converge on the position that 'there is no gold standard' (Cartwright, 2007, abstract). Or as Deaton (Deaton, 2010, p.424) says, 'gold standard' thinking is mistaken because 'experiments have no special ability to produce more credible knowledge than other methods.' This has had some effect on practice. Just as the demand for evidence-based policy has largely softened to a call for evidence-informed policy in light of the other factors that are legitimately relevant to policy decisions, so the calls for evidence of 'what works' have softened to often include the realist inflection 'for whom, in what circumstances' (Pawson and Tilley, 1997, p.220). Despite this shift in tone, unfortunately such avowals of the importance of context are often purely theoretical, and not reflected in practice by researchers or development experts. For example, Howard White, long-time director of 3ie, has written at length on the importance of paying attention to context (White, 2009, 2010). Nevertheless, 3ie continues to produce systematic reviews and maps of evidence that do not provide even basic information about the contexts in which effect sizes have been measured, let alone attempt to use such information to inform the appraisal.

Gold standard thinking and gold standard practice persist despite injunctions against them from philosophers of science, non-economists, and even some senior economists. The effect on qualitative approaches to research is particularly marked, as they are 'seen as failing to provide hard, reliable, factual data' (Sanderson, 2000, p.436). The tendency to describe the systematic review as a tool for aggregating 'current global evidence' (DFID, 2011, p.i), is also worrying. This erases all evidence that falls outside of the inclusion criteria for the review, even where such evidence was key to the development of the programmes being reviewed. Excellent examples of this can be found in the conditional cash transfers literature. Qualitative evidence was essential for the development of Mexico's PROGRESSA/Oportunidades programme and Nicaragua's PRS programme, allowing those programmes to be tailored to their local environments in ways that contributed to the size of their effects (Adato, Hoddinott and Emmanuel, 2010; Levy, 2007). Despite this, recent reviews of the evidence have focussed on aggregating mean treatment effect estimates of different transfer programmes in different contexts to create supposedly globally valid estimates of the effects of future programmes (DFID, 2011; Fiszbein and Schady, 2009; Garcia and Saavedra, 2013).

Observational quantitative methods also suffer from 'gold standard' thinking. Deaton (2010) describes a growing distrust of econometric analysis and any other non-experimental evidence that has created a climate in which development is increasingly seen as a process that occurs at the micro level, rather than a process of structural transformation of the economy. Chang (2011) calls this '*ersatz* development' and believes it to be an approach to explaining development that is as conceptually lacking as an approach to explaining Hamlet without reference to the prince of Denmark. Pritchett and Kenny (2013) criticise development economics' increasing focus on the micro level as 'kinky development', for its preoccupation with putting a kink in the distribution of consumption rather than shifting the entire distribution far to the right.[4]

These pernicious effects of the tendency to think uniquely in terms of internal validity when judging evidence quality based on method choice suggest that a more powerful approach is required than merely explaining why 'there is no gold standard'. The uniquely negative project of philosophers and methodologists to date in attacking gold standard thinking may have lacked credibility, as saying that 'there is no gold standard' without endorsing a positive project seems worryingly close to arguing that 'anything goes'. While Cartwright concludes the abstract to her

[4] Though, interestingly, Kenny (2021), another researcher at the Center for Global Development has suggested that the micro-driven focus on 'kinky development' may be right for the wrong reasons. Kenny suggests that we should be focussed on kinking the tail of the global income distribution because the declining marginal utility of income means that returns to development investments in middle-income countries need to be expected to be up to 16x higher than investments in the poorest countries in order to be equally valuable.

2007 article with ‘[t]here is no gold standard’, it is more positively offered that ‘which method is best depends case-by-case on what background knowledge we have or can come to obtain.’ However, to a development expert or a policy maker seeking advice about which evidence to trust when assessing the effectiveness of development interventions, this sort of statement is equivalent to ‘it depends’. Such advice offers no guide for further action and doesn’t chime with the reasonable intuition that for certain sorts of question, some methods are *in general* more reliable than others. It is the contention of this research project that what is needed is a positive account of the relative merits of evidence generated using different methods that considers ability to generalise from results as well as internal validity. The originality of this research project inheres in its attempt to address the lack of such a positive project. Its significance lies in the possibility of satisfying the demand for such an account.

2.5 LIMITING THE SCOPE TO (QUASI-)EXPERIMENTAL IMPACT EVALUATION

2.5.1 Limiting the scope to effectiveness

As the previous section has shown, ‘gold standard’ thinking has damaging effects on many types of evidence by suggesting that any evidence that does not conform to the gold standard is less valuable or even without value. However, purely negative accounts of the problems with such thinking have not been persuasive, and a systematic, transparent account of the relative merits of evidence issued from different methods is needed to show that abandoning gold standard thinking does not mean surrendering to ‘anything goes.’ Petticrew and Roberts (2003) famously went beyond a purely negative account of gold standard thinking to acknowledge that there are ‘horses for courses’, and therefore that some methods are in general stronger than others when addressing particular sorts of question. In trying to develop a positive account of evidence quality that goes beyond interval validity to also consider the transferability of results, I must limit myself to one of these courses and assess horses relative to it. A systematic account of evidence quality *in general* is well outside the scope of a thesis and in any case may not be a coherent project. Consider three types of research question adapted from Petticrew and Roberts (2003, p.528): 1) Salience – does the issue matter? Does the intervention respond to people’s concerns? 2) Process of delivery – how was the intervention adapted and deployed in this particular context? 3) Effectiveness – to what extent did the intervention cause changes in outcomes of interest for this population? Different methods deployed in different ways will be more suitable to respond to each of these sorts of questions. As a result, methods to generate evidence in response to each of these questions will have different markers of quality.

In this research project, I limit myself to assessing the quality of evidence designed to respond to the third type of question, concerning effectiveness. As discussed, a sufficient reason to bound the scope of this research to evidence that addresses one type of question is that to try to do more would be impossibly hard in the time available and in any case possibly incoherent. I choose to examine the quality of evidence on effectiveness questions because the most reasonable form of gold standard thinking acknowledges the existence of different types of evidence for different types of research question but maintains the supremacy of randomised controlled trials for answering questions of effectiveness (Oakes, 2018). This limited and more reasonable form of gold standard thinking is nonetheless based purely on internal validity concerns and is undermined by the problem of external validity discussed in Section Three. It is therefore productive to attempt to replace it with a more general account of the relative merits of evidence generated using different methods for addressing questions of effectiveness. At this stage, a candidate question emerges for the primary research question of this research project:

> *On questions of the effectiveness of development policy interventions, can we give a useful, systematic account of the relative merits of evidence generated using different methods that goes beyond internal validity to also consider external validity?*

2.5.2 Limiting the scope to large-*N* studies

The candidate research question above remains too broad in scope to be addressed by this research project. Furthermore, it still does not target the provision of an alternative to the most reasonable formulation of the gold standard hypothesis. This is because there are two broad approaches to effectiveness questions that are possible depending on the total number of units of assignment for the intervention being studied, *n*. Units of assignment may be individuals or something else, such as villages, households, counties or even countries. If *n* is sufficiently large, quantitative analysis using statistical techniques is possible. If *n* is not sufficiently large, then such methods are not possible. Quantitative approaches such as the synthetic control method or economic modelling may still be appropriate, but not the use of statistical techniques (White, 2010). Therefore, the most reasonable form for gold standard thinking is that randomised controlled trials are *the best* method for assessing the *effectiveness* of interventions that are assigned over *a large number of units*. This research project attempts to provide an alternative to that account of evidence quality. In doing so, I consider the universe of desirable methods for investigating such questions to extend only to ‘impact evaluations’ in the sense identified by White (2010, p.154).

> *Impact is defined as the difference in the indicator of interest (Y) with the intervention (Y_1) and without the intervention (Y_0). That is, impact = Y_1 - Y_0 (e.g. Ravallion, 2008). An impact evaluation is a study which tackles the issue of attribution by identifying the counterfactual value of Y (Y_0) in a rigorous manner.*

What is meant by ‘rigour’ here? The following paragraphs explain.

Questions of intervention effectiveness are causal questions. Answering them requires overcoming the attribution problem; we must separate the effects of the intervention being studied from the confounding effects of other causal processes that might also be responsible for changes in outcomes of interest. Therefore, simple comparisons of the levels of outcomes of interest for recipients of an intervention before and after the introduction of the intervention cannot overcome the attribution problem. In order to do so, some kind of comparison with non-recipients of the intervention is required. This sort of comparison is always possible. Even where entire regions or countries have been exposed to the same intervention, comparison with other regions or countries can provide a relevant comparison to attempt to overcome the attribution problem (Card and Krueger, 1993).

Once some comparison group of non-recipients of the intervention has been identified, a further barrier to reliable causal analysis presents itself. It is possible that recipients and non-recipients of an intervention are systematically different from each other. This might be because recipients chose or were chosen to receive the intervention on the basis of some characteristic(s). These characteristics might causally influence outcomes of interest, resulting in biased estimates of the intervention's effect if these characteristics are not controlled for in some way. Pritchett and Sanderfur (2015) are quite right that this selection bias may be less important than other biases, such as the bias arising from comparisons across contexts. However, there is no reason not to minimise selection bias. Therefore, some kind of matching of recipients with non-recipients of the intervention on causally relevant characteristics should always be attempted when addressing effectiveness questions.

What we would like to know, when identifying the effectiveness of an intervention, is what the difference in outcomes of interest is between what did happen to those outcomes for recipients of the intervention and what would have happened to those same outcomes for those same people had the intervention not taken place (Lewis, 2001). This later situation is, by definition, a counterfactual reality that did not come to pass. For this reason, it is never possible to examine it. As Section Two has described, the power of experimental and quasi-experimental approaches is that they minimise the assumptions required to consider a constructed comparison group equivalent to the counterfactual group.[5] This minimising of assumptions is what is meant by 'rigour' in the definition of an impact evaluation from White, above. While experimental methods cannot always be employed in an assessment of the effectiveness of a large-*N* intervention, some form of matching can always be employed. As matching methods are impact evaluation methods, an impact evaluation method is always possible for large-*N* evaluations of effectiveness.

So, the scope of this research project is limited to the set of methods that answer *attribution* questions in assessing the *effectiveness* of interventions delivered to *a large number* of treatment

[5] Random allocation to treatment is often spoken about as if it ensures that no further assumptions are required to treat the comparison group as equivalent to the counterfactual. However, the possibility of non-random differences in post-randomisation changes as well as of 'unhappy' randomisation means that further assumptions must always be argued for (Deaton and Cartwright, 2018b).

units, for which it is necessary to estimate the *counterfactual*. For White and many others[6] the term ‘impact evaluation’ only applies to the experimental and quasi-experimental (including matching) methods that are designed to answer this specific sort of question. However, this restrictive definition of ‘impact evaluation’ is contested. Several influential reports have used a much broader definition including methods seeking to address a broader set of possible evaluation questions including that from the Network of Networks on Impact Evaluation (Leeuw and Vaessen, 2009) for the World Bank and Stern et al. (2012) for DFID. In this thesis I attempt to be clear by prefacing ‘impact evaluation’ with ‘(quasi-)experimental’ and remind readers where this is important that I am addressing questions of method choice only in the context that the evaluation questions of interest are attribution questions concerning effectiveness for an intervention with a large *N*.

In light of the limiting of scope argued for in this section, the primary research question that guides this project becomes:

> *Can we give a useful, systematic account of the relative merits of evidence generated using different (quasi-)experimental impact evaluation methods that goes beyond internal validity to also consider external validity?*

The much-reduced scope of this research question nonetheless permits contributions to debates about evidence quality outside of the (quasi-)experimental impact evaluation literature. What is learned in the assessment of the quality of (quasi-)experimental impact evaluations can be applied more widely to ameliorate some of the damaging consequences of the more generalised versions of gold standard thinking discussed earlier in this chapter. This is discussed further in Chapter Nine. In the next chapter, the primary research question will be given a theoretically rich interpretation that overcomes a critical ambiguity of ‘external validity’ discussed in that chapter. The primary research question will then have reached its final form. However, further examination will motivate splitting the primary research question into two research

[6] See, for example, Gertler et al.’s (2016) extensive and influential textbook.

subquestions that address separate aspects of it and can each be given a yet more theoretically rich interpretation.

3 Theoretical approach

The review of the literature conducted in the previous chapter motivates a candidate primary research question:

> *Can we give a useful, systematic account of the relative merits of evidence generated using different (quasi-)experimental development impact evaluation methods that goes beyond internal validity to also consider external validity?*

This research question contains some implicit assumptions supported by the literature review of the previous chapter, which are restated here for clarity. The scope of this research is limited to the standards of (quasi-)experimental impact evaluation quality that are applied during the research to policy process. It is not naïvely assumed that development policy is or should be 'evidence-based'. However, it is a starting assumption of this research that evidence about the effectiveness of policy interventions aiming to promote development does, at least sometimes, play a role in the formulation of development policy. Further, it is a starting assumption of this research that standards of (quasi-)experimental impact evaluation quality active in communities of development 'experts' act as filters on the sorts of evidence that make their way into policy.

This chapter moves iteratively between the research question and the theoretical literature to give a theoretically rich interpretation to the primary research question and to arrive at a research strategy for answering it. In Section One, it is argued that 'external validity' is too ambiguous to provide a framing for this research project. Different intended meanings of 'external validity' are identified in the literature, and an argument is presented that 'transferability' is the appropriate framing for this research project. In Section Two, the primary research question is divided into two research subquestions that address two complementary aspects of the primary research question. The rest of Section Two consults the theoretical literature to identify an ontological account and epistemic strategy that is capable of underpinning an answer to the first research subquestion. The first research subquestion is then given a theoretically rich interpretation using the terminology of the account identified, scientific realism. Section Three builds an argument for the suitability of the 'critical turn' in scientific realism as a way of understanding the necessity of crafting a 'useful' answer to the primary research question. It is then argued that an ontological account including 'epistemic communities' is suitable to integrate with a critical realist approach to provide a research strategy for answering subquestion two.

3.1 FROM 'EXTERNAL VALIDITY' TO 'TRANSFERABILITY'

3.1.1 What's wrong with 'external validity'?

It might be argued that the correct framing for a response to subquestion one is a framing in terms of external validity. Chapter Two, Section Three motivated the primary research question by talking about a 'problem of external validity'. Why not continue with this framing? Initially, this research project did progress using that framing, but the interpretation of results uncovered a conceptual confusion which prompted further investigation of the literature and led to the understanding that a framing in terms of external validity reduced clarity by being ambiguous between several alternative meanings.

Confusion around the meaning of 'external validity' is widespread in the social science evaluation literature and high-profile. In response to Deaton and Cartwright's (2018b) instantly seminal discussion of the power and limitations of the RCT method in social science, Imbens (2018) takes those authors to task for their use of 'non-standard' definitions of 'internal and external validity'.[7] However, the apparently canonical definitions offered by Imbens are unclear and incompatible with each other. Imbens (2018, p.51) writes:

> *By the standard usage I mean, for example, Shadish et al. (2002) who define internal validity as "the validity of inferences about whether observed covariation*

[7] This discussion takes place in a special edition of *Social Science and Medicine* dedicated to Deaton and Cartwright's paper, many responses to the paper from leading figures in the social science impact evaluation literature, and Deaton and Cartwright's response to those responses. The usefulness of the research project described in this thesis is suggested by the fact that it addresses several of the issues debated and brings a novel empirical approach to bear on them.

> *… reflects a causal relationship," and external validity as "the validity of inferences about whether the cause-effect relationship holds over variation in persons, settings, treatment variables, and measurement variables.["] Rosenbaum (2002) writes in a similar spirit, "A randomized study is said to have a high level of 'internal validity' in the sense that the randomization provides a strong or 'reasoned' basis for inference about the effects of the treatment … on the … individuals in the experiment," and " 'external' validity refers to the effects of the treatment on people not included in the experiment."*

Clearly, these two proposed definitions are not the same as each other. A) 'the validity of inferences about whether the cause-effect relationship holds over variation in persons, settings, treatment variables and measurement variables' is not equivalent to B) the validity of inferences about 'the effects of the treatment on people not included in the experiment'.

Definition B) is more confusing than A). If 'the treatment' in B) is taken to refer to the particular instance of the treatment evaluated, bounded by the experimental context, then 'the effects … on people not included in the experiment' might be taken to mean the unobserved spillover effects of that treatment. This would be a very non-standard use of 'external validity' indeed and would be obviously incompatible with A). More charitably, if 'the treatment' is taken to refer to an intervention that was deployed in some wider context Φ of which the experimental context φ is a subset, then 'the effects … on people not included in the experiment' might be taken to refer to the treatment effects in Φ. The validity of inferences about treatment effects in this wider population based on observations of treatment effects in its subset φ will be a judgement of the representativity of φ as a sample of Φ including a judgement of the consistency of implementation of 'the treatment' between the two contexts. So, in light of the clarificatory work above and in order to adopt consistent terminology throughout this research project, we can re-express B) as 'the validity of inferences about the extent to which treatment effects from some sample context will hold in the population context'.

Definition A) does not suggest that ‘external validity’ in the sense that it defines is limited to the validity of inferences about populations based on samples. It is a more general, more ambitious definition that seeks to capture the extent to which inferences about causal relationships between intervention, context and outcomes are valid for any context θ that is not causally equivalent to the study context φ. As was discussed in Chapter Two, Section Three, this includes a non-random sample of φ or φ itself at a different time. The case of generalising from study context φ to its parent context Φ, of which it is a subset, is a special case of the more general process of attempting to use results from φ to make inferences about some other context θ. Shadish *et al.* (2002, p.22), quoted by Imbens above, support this interpretation of their definition when they say that ‘[w]hether from narrow to broad, broad to narrow or across units at about the same level of aggregation, all these examples of external validity questions share the same need – to infer the extent to which the effect holds over variations in persons, settings, treatments, or outcomes.’ So, in light of the clarificatory work above and in order to adopt consistent terminology throughout this research project, we can re-express A) as ‘the validity of inferences about the extent to which treatment effects will hold in some other context’. Definitions A) and B), then, are not equivalent, but one can be thought of as a special case of the other.

Though B) seems equivalent to ‘the validity of generalising inferences’ or ‘generalisability’, it is unclear what term should be used to refer to A). One could use ‘external validity’ to refer to A) and see generalisability as a special case of external validity. In this way, ‘external validity’ would remain the umbrella term and special instances of it could be referred to by other terms. Perhaps, then, this research project could be framed in terms of external validity after all. However, as the next paragraph will demonstrate, there is a further sense of ‘external validity’ that has some support in the social sciences evaluation literature and the term is therefore too contested and ambiguous to be sufficiently clear.

As discussed in Chapter Two, Vivalt (2016, 2019) uses a large set of (quasi-)experimental impact evaluations to assess the observed heterogeneity of treatment effects across studies of the same class of intervention targeting the same outcome. Vivalt (*ibid*) calls this an assessment of external validity. For Vivalt, external validity concerns are concerns about the correspondence between study treatment effect estimates and the 'true' treatment effect across all contexts. As the 'true' treatment effect cannot be observed, we are compelled to compare results with the closest approximation we can observe, the average treatment effect across studies, perhaps modified by some interaction terms with moderator variables to take account of the specific context in which the study or studies have been conducted. Call this definition C) 'the extent to which given treatment effects are predictive of the mean treatment effect across all observed contexts.' This definition implies that external validity can only be known *ex post* and cannot be established *ex ante* with reference to any specific target context. This way of using 'external validity' pervades the development economics literature. For example, it is the meaning implied by Angelucci et al. (2010, p.214) when they say that '[t]o provide external validity to the constructed data on extended family links in the PROGRESA data, we present similar information from an alternative data set that was collected in a comparable economic environment and time period.'

3.1.2 Three alternative terms for three intended meanings of 'external validity'

As Cartwright and Deaton (Deaton and Cartwright, 2018a, p.87) acknowledge in their reply to Imbens, 'it is impossible to change the use of the terms "internal validity" and "external validity"'. A more promising route, then, seems to be to accept the ambiguity of 'external validity' and to attempt to define new technical terms for each of its possible meanings. This subsection proposes some technical terms to capture the different uses or intended meanings of 'external validity' observed in the literature and argues that this research project should be framed in terms of 'transferability'. For ease of reference, the observed uses or meanings of 'external validity' are summarised in Table 3.1, along with the suggested terms for each meaning. Next, the argument for the suitability of each term is presented.

Table 3.1: Meanings of 'external validity' with alternative suggested specific terms

	Meaning	**Example**	***ex-ante* /** ***ex-post***	**Suggested specific term**
A)	The extent to which treatment effects will hold in some other context	'external validity (whether valid inferences are drawn for other projects, either as scaled up versions of that project in the same setting or as similar projects in different settings)' (Ravallion, 2009, p.32)	*ex-ante*	Transferability
B)	The extent to which treatment effects from some sample context will hold in the population context	'external validity—the relevance of the IE [impact evaluation] to the scale-up of the programme in a given country' (Davis et al., 2016, p.65)	*ex-ante*	Generalisability
C)	The extent to which given treatment effects are predictive of the treatment effect across all observed contexts	'the "external validity" of one set of *n* results … *i.e.* how well observed point estimates $Y_1, Y_2, \ldots, Y_n$ can be used to jointly predict the point estimate of another study, Y_j' (Vivalt, 2019, p.12)	*ex-post*	Observed heterogeneity

Taking the meanings of 'external validity' identified in the previous subsection in reverse order, the first meaning in need of a specific technical term is C). As Vivalt herself suggests, a good fit for this meaning is the 'observed heterogeneity' of treatment effects.

Moving on to B), because we are talking about making inferences about a superset based on observations of a subset, the most apt term in this case appears to be 'generalisability'. In natural language, 'generalise' is used in this way, as the act of 'mak[ing] a general or broad statement by inferring from specific cases' (Soanes and Stevenson, 2004). Confusingly, some authors within the evaluation literature have sought to define technical senses of 'generalise' that include reasoning from the general to the particular or across cases at the same level of generality, but these are relatively few.[8] Many others have treated ‘generalisability’ as an ambiguous stand-in for external validity. However, still other attempts to codify terms in the evaluation literature have defined ‘generalisability’ similarly to B), including recent high-profile efforts such as Deaton and Cartwright (2018b) and systematic approaches to reviewing the literature such as Davey et al. (2018). In the absence of a more promising candidate, and in the hope that the natural language similarity is a positive predictor of a technical term’s likelihood to establish a hegemonic definition, ‘generalisability’ seems the best candidate to express the meaning sometimes associated with ‘external validity’ captured by the definition B).

The previous paragraph argued that ‘generalisability’ is so close in its natural language meaning that it seems the clear choice for a technical term defined as in B), but there is no obvious natural language candidate for A). The two leading candidates for terms to refer to A) are ‘transferability’ and ‘transportability’. ‘Transferability’ is defined by Burchett, Umoquit and Dobrow (2011, p.239) as ‘[t]he likelihood that the study’s findings could be replicated in a new, specific setting (i.e. that its effectiveness would remain the same[.])’ This may seem like a very restrictive definition. As Deaton and Cartwright (2018b, p.10) point out, *‘simple extrapolation’* based on the belief that the result holds everywhere is excessively naïve, making a definition of transferability in these terms too demanding to satisfy. However, by framing their definition probabilistically, Burchett et al. (2011) give a definition that is equivalent to A) on the undemanding assumption that the extent to which a result might be expected to hold is interpreted as equivalent to the likelihood that it holds. In a similar spirit, Pearl and Bareinboim (2014, p.579) define a ‘problem of transportability’ as the conditions under which we are ‘license[d] to transfer causal effects learned in experimental studies to a new population’.

So ‘transferability’ or ‘transportability’ are both good candidate terms for A).[9] However, ‘transport’ also has an established alternative meaning in the evaluation literature, referring to

[8] See, for example, Shadish et al. (2002), Sculpher et al. (2004)

[9] Davey et al. (2018) detect a distinction between ‘transportability’ in Pearl and Bareinboim’s usage and ‘transferability’ in Burchett et al.’s. According to Davey et al. (2018), ‘transportability’ is intended to be more broad than transportability, referring for some target context to the ‘use of any causal knowledge and not only to whether or not the size of the effect is likely to be the same.’ However, I cannot discern

the movement of an intervention (rather than its results) from one setting to another (Leijten et al., 2016). Further, ‘transport’ emphasises a binary judgement of the transferability of results. This is true both by analogy to its natural language interpretation – transport either happens or it doesn’t, whereas arguably something can be transferred to an extent – and in its usage by Pearl and Bareinboim, who seek to define the conditions under which a result can be transported (wholesale). Therefore, ‘transferability’ seems the more promising candidate term for a specific technical term to refer to meaning A).

3.1.3 An argument for framing the primary research question in terms of ‘transferability’

The question remains, in terms of which of the three technical terms now defined should this research project be framed? As Chapter Two has described, this research project is motivated by the observation that we have no systematic approach to appraising the utility of (quasi-)experimental impact evaluation results from different methods for overcoming the ‘problem of external validity’ discussed in Chapter Two, Section Three. Let us consider a possible framing in terms of C). The observed heterogeneity of treatment effects is not the cause of nor is it a sufficient condition for the ‘problem of external validity’. However, it is a necessary condition for treating that problem urgently. If results of similar programmes across differing context were in fact very similar, then it would not be so problematic that we require challenging arguments to justify using results from a (quasi-)experimental impact evaluation in one context to predict results of a similar program in a different target context. The observed heterogeneity of treatment effects means that this problem is, in fact, urgent. However, no amount of discussion of the observed heterogeneity of results across contexts will provide premises for an argument in answer to the primary research question. This is because answering that question hinges on investigating the possibility of providing an account of (quasi-)experimental impact evaluation quality that considers the facilitation of *ex ante* judgements of the utility of results from the study context to inform thinking about some target context. Because the need to do this discussed in Chapter Two is not limited to instances of generalisation from a sample study

this difference in usage when examining Pearl and Bareinboim’s work, which deals with the transport of ‘results’ and ‘causal effects’ that seem in the examples to be equivalent to treatment effects.

context to a target population context, it is in terms of transferability that this research project must be framed.

Many of the arguments for 'external validity' made for the findings of (quasi-)experimental development impact evaluations are made in terms of B). These arguments attempt to overcome the 'problem of external validity' set out in Chapter Two without surrendering to Cartwright's demand to provide a model of intervention causation in combination with causally relevant contextual factors. Instead, interventions are conceptualised as 'products' that replication in many different contexts can 'accredit as effective' (Bonell et al., 2012, p.2300). This process of accreditation implicitly assumes that the contexts in which the intervention is evaluated can be argued to be a representative sample from a total population of contexts in which the intervention might be implemented. As Banerjee and Duflo (Banerjee and Duflo, 2008, p.16) put it: '[i]f we were prepared to carry out enough experiments in varied enough locations, we could learn as much as we want to know about the distribution of the treatment effects across sites conditional on any given set of covariates.' The problem with this approach is that without a theory of how the intervention works and the causally relevant contextual factors that mediate its effectiveness, we cannot provide a compelling argument that the contexts in which it has been evaluated are representative of the total population of possible implementation contexts. As Banerjee and Duflo themselves admit, in the absence of such a theory 'we should ideally choose random locations within the relevant domain' (*ibid*, p.14). This amounts to a call for the external to be rendered internal. If researchers must randomly sample clusters of individuals from 'the relevant domain' and then run (quasi-)experimental impact evaluations on all of those clusters, then they must in effect create an enormous sampling frame for a test population comprised of all the individuals in the world (present and future) that the intervention could potentially benefit. This is clearly an impossible task, as Pawson and Tilley (1997, especially p.118) and Cartwright (2007, 2008) persuasively argue. In the special case of generalisation from a (quasi-)experimental impact evaluation conducted in a representative sample of the relevant context(s), an argument for generalisability can be made. Otherwise, the only way to overcome the 'problem of external validity' identified in Chapter Two, is to make an argument in terms of transferability. It is therefore in terms of transferability that the primary research question must be restated:

> *Can we give a useful, systematic account of the relative merits of evidence generated using different (quasi-)experimental development impact evaluation methods that goes beyond internal validity to also consider the transfer of results to other contexts? If so, how?*

This question is adequate but can be given a final tweak to reflect an important consideration highlighted by Deaton and Cartwright (2018a) in their reply to Imbens (2018), discussed above. While Deaton and Cartwright do not seek to reject Imbens' proposed definitions of external validity, they take issue with one implication that they consider to be presupposed by those definitions. They worry that those definitions presuppose that 'external validity' is taken to be a property of an estimate or even of a study design, independent of other facts about the world. In order to be absolutely clear that this is not presupposed in this research project, the primary research question is reformulated as follows:

> *Can we give a useful, systematic account of the relative merits of evidence generated using different (quasi-)experimental development impact evaluation methods that goes beyond internal validity to also consider the extent to which methods facilitate the transfer of results to other contexts? If so, how?*

The use of 'facilitate' here is intended to convey an acceptance of the fact that the transferability of results or treatment effects is dependent on facts about the world external to the evaluation in question. The best that an evaluation design can do is to provide *some of* the necessary premises for an argument for the transferability of results to some other context. Estimates of treatment effects are not transferable from study context φ to target context θ because the evaluation was designed well, but rather they may be transferable (or not) based on the similarities (or differences) in the causal structures of φ and θ. A well-designed evaluation using a high-quality method will facilitate such an argument to a high degree, but that does not make transferability a property of the design or of the estimate independent from other facts about the world.

3.2 ANSWERING RESEARCH SUBQUESTION ONE: A REALIST RESEARCH STRATEGY

This section and the next consult the literature to craft a theoretically informed research strategy for answering the primary research question. The first step in this process is to split that question into two subquestions, that occurs in the first subsection of this section. Then, in the second and third subsections of this section I assess the theoretical literature to identify an epistemic strategy for answering the first research subquestion. Section 3.3 will do the same for the second research subquestion.

3.2.1 Separating questions of 'usefulness'

The primary research question given at the end of the previous section asks how we can give a *useful, systematic* account of the relative merits of evidence generated using different (quasi-)experimental impact evaluation methods that goes beyond internal validity to also consider the transferability of results. Answering that research question therefore requires two distinct but complementary strands of research. On the one hand, a positive account must be developed of the relative strength of (quasi-)experimental impact evaluation methods with reference to the transferability of their findings. On the other hand, an understanding must be developed of how this account can be useful. This motivates the framing of two subquestions, the answers to which will be combined to answer the primary research question.

1. *What are the systematic differences, if any, between (quasi-)experimental impact evaluation methods regarding the extent to which they facilitate the transfer of their results to other contexts?*

2. *What features would an account of these differences need to have in order to be useful to development experts?*

3.2.2 Why adopt a realist approach?

The approach to thinking about causation implicit in the concept of 'external validity' as it is used in C) and as it is often used in technical economics literature is the successionist theory of causation, in which causation is reducible to covariation, with causes having the property of temporally preceding their effects. This way of thinking about causation is anti-realist about causal powers or ‘mechanisms;’ implicit in it is the judgement that causal powers are not legitimate objects for scientific study. Harré (1985, p.116) identified the successionist theory of causation as one of ‘the two great metaphysical theories of causality’ in the following passage:

> *‘In the* generative *theory the cause is supposed to have the power to generate the effect and is connected to it. In the* successionist *theory a cause is just what usually comes before an event or state, and which comes to be called its cause because we acquire a psychological propensity to expect that kind of effect after the cause.’*
>
> (*ibid*)

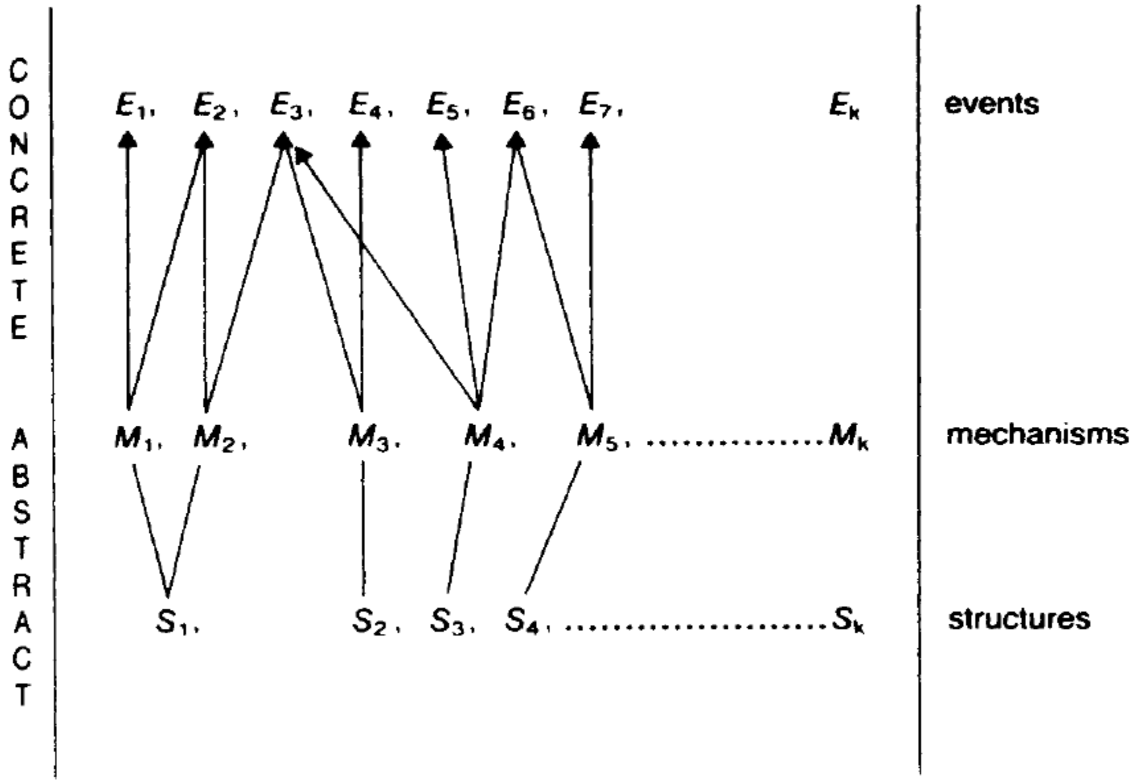
The successionist theory of causation could be objected to on ontological grounds, as do many philosophers of science (Harré, 1970, 1985; Bhaskar, 1975; Sayer, 1992). However, for the purposes of this research it is sufficient to observe that the successionist theory of causation cannot provide for *ex ante* assessments of transferability. As discussed in Chapter Two, Section Three, such assessments would have to be based on a discussion of how intervention T is supposed to have the power to generate outcomes O and an argument about the extent of the relevant similarities between the causal structure of the study population φ and the target population θ. This calls for a generative approach to causation. As realism is the ontological framework and associated epistemic strategy that underpins generative accounts of causation, realism is the theoretical framework within which this research project must be situated (Sayer, 1992).

3.2.3 Crafting a realist research strategy

To understand how research subquestion one might be operationalised to permit its investigation through research, it is necessary to unpack the realist understanding of causation a little more. The realist ontological framework divides reality into three levels: events, mechanisms, and structures (*ibid*). Events are caused by mechanisms, which overlap with each other to magnify or frustrate each other’s action, creating complex open systems of causation

(Bonell et al., 2012). These mechanisms are emergent properties of underlying structures. We have different levels of access to each of these layers of reality. Events can be directly observed, whereas the action of mechanisms can only sometimes be directly observed, and the underlying structure of reality is hardest of all to gain access to. For some explanations of some phenomena, such as explanations of physical processes, some level of access to mechanisms and structures through observation might be possible. For others, such as explanations of some social processes, only entities can be observed, and the existence of mechanisms and structures must be inferred (Bhaskar, 1975). Nevertheless, in all cases the three levels of reality are legitimate objects of scientific study, though some are more 'concrete' and others more 'abstract' (Sayer, 1992, p.117). Sayer's (*ibid*) Figure 8 is reproduced below as Figure 3.1, to illustrate realist ontology.

Figure 3.1


CONCRETE
ABSTRACT
E_1, E_2, E_3, E_4, E_5, E_6, E_7, E_k
events
M_1, M_2, M_3, M_4, M_5, ... M_k
mechanisms
S_1, S_2, S_3, S_4, ... S_k
structures

(Sayer, 1992, p.117: Figure 8)

An answer to research subquestion one could be developed by using realist ontology to examine the extent to which different research methods provide for different levels of exploration and reporting of the causal features relevant to arguments for the generalisability of findings. The causal features of interventions must be investigated anew for each context because, as Pawson *et al.* observe, 'the "same" intervention never gets implemented identically' (Pawson et al., 2004, p.v). The distinction between contextual mechanisms and intervention mechanisms is central to Pawson and Tilley's (1997) influential realist approach to the evaluation of policy interventions. In their framework, interventions are conceptualised as embodiments of some theory of the form 'if we do X in this way, then it will bring about an improved outcome' (Rycroft-Malone et al., 2012, p.2). Evaluations of interventions should seek to test and refine this programme theory so as to build an ever more robust theory of the effects of intervention mechanisms, enabled or frustrated by contextual mechanisms, on outcomes. With the evaluation of each new intervention, programme theory is continually developed, tested, and refined, through realism's distinctive retroductive epistemic strategy. The hope of realist evaluators is that in this way, the body of knowledge of context-mechanism-outcome combinations should grow larger and more robust so as to build an ever more 'practically adequate' picture of how causation works in the world (Sayer, 1992; Pawson and Tilley, 1997, especially p.220; Sanderson, 2000).

Bhaskar (1975) advocated the use and interpretation of experiments by realists. However, unfortunately, many social scientist realists, following Pawson, do not believe that the findings of experimental studies are ever intelligible through a realist lens. According to these researchers, this is because experiments are 'fundamentally built upon a positivist ontological and epistemic position', that fails to take account of the complexity of social causation by 'merely controlling for it' (Marchal et al., 2013, pp.124–125). Rather than making suggestions for how random allocation to treatment can be used to test and develop rich causal theories, such realists have preferred to dismiss experimental evidence entirely and focus on developing alternative evaluation methods such as Realist Evaluation (Pawson and Tilley, 1997). This is particularly surprising because such realists make extensive use of natural experiments to support their arguments for counterfactuals (*ibid*). Bonell et al. (2012, 2013, 2016) have persuasively argued that trials can be realist in the strong sense that they are compatible with

Realist Evaluation and can be used to test hypotheses about context-mechanism-outcome configurations. In more recent work, Bonell et al. (2018) also point out that many trials are already realist in the less demanding sense, in that they are based on the assumption that causal mechanisms are legitimate objects of study. *Contra* Pawson, and in agreement with Bonell et al., it is the contention of this research programme that random allocation to treatment does not make evaluation results illegitimate or unintelligible for realists. I argue that the ontological framework and retroductive research strategy of realism can provide a means of assessing the extent to which any method facilitates the generalisation of findings. A method for doing so is detailed in Chapter Four.

As discussed in Chapter Two, Section Three, an argument for the application of an evaluation's findings from study population φ to some target population θ requires a model of intervention causation, and models of the causal structure of φ and θ. The question of the relative strength of (quasi-)experimental impact evaluation methods with respect to transferability, then, becomes a question about the extent to which different methods explore and report on intervention causation and the causal structure of φ. The causal structure of φ can be further unpacked, following the realist analysis of intervention causation outlined above, into intervention mechanisms and those contextual mechanisms that interact with intervention mechanisms, either to frustrate or enable them. It is conventional in realist evaluation practice to refer to intervention mechanisms as 'mechanisms' and to contextual mechanisms which act as barriers and/or enablers to intervention mechanisms as 'context' (Pawson and Tilley, 1997; Westhorp, 2014). However, in choosing terminology for the primary research question I seek to distinguish between those features of context that are causally relevant to treatment effects and those that are not. Therefore, intervention mechanisms are referred to as 'intervention mechanisms' or just 'mechanisms.' The contextual features that are relevant to intervention causation are referred to as barriers and enablers.

It might be objected that the terms 'barrier' and 'enabler' should be reserved specifically for binary conditions that are either present or not. After all, a barrier is 'a fence or other obstacle that prevents movement or access' (Soanes and Stevenson, 2004). In this telling, the absence of barriers and the presence of enablers perhaps correspond specifically to Mackie's (1965) INUS and SUIN conditions. However, as anyone who has climbed, gone around, or even stepped over a barrier knows, it is equally plausible to use 'barrier' to refer to something that, to some degree, *frustrates* movement or access. Development practitioners use the term 'barrier' in this second way as well as in the first. For example, Galasso (2006, p.6) gives high transaction costs associated with application processes as an example of a 'barrier' which frustrates take-up of social programs to different extents for different individuals. Similarly, an 'enabler' is sometimes thought to be uniquely necessary for the action of an intervention, jointly necessary, or sometimes merely helpful, magnifying the effect. In this thesis I use the terms 'barrier' and 'enabler' in the most general sense, to refer to states of affairs that may be hard stops or merely frustrating, necessary or merely helpful. In this usage these terms are intended to capture the full range of factors that moderate the effect of an intervention in a context.

Perhaps now we can give a theoretically rich, realist interpretation of research subquestion one? We might try:

> *What are there systematic differences, if any, between (quasi-)experimental impact evaluation methods regarding the extent to which they explore and report on the barriers and enablers of intervention mechanisms present in the study context?*

Unfortunately, reporting the barriers and enablers of intervention mechanisms present in the study context is not sufficient for an argument for the transferability of results to some target context where the intervention acts through more than one mechanism. If intervention T acts through one mechanism M, mediated by contextual barriers and enablers C_{M1}, C_{M2}, … , C_{Mn}, then all that is required for an argument for the transferability of treatment effect O from study context φ to target context θ is that the evaluation of T in φ report C_{M1}, C_{M2}, … , C_{Mn}. However, consider the case where intervention T acts through two mechanisms, L and M. It is not sufficient for an argument for the transferability of treatment effect O merely to report C_{M1}, C_{M2}, … , C_{Mn} and C_{L1}, C_{L2}, … , C_{Ln} without some way of assessing the relative importance of those two sets of contextual barriers and enablers. What is required is knowledge of the relative importance of M and L in causing O in *φ*. In order to assess this, the evaluation must generate and report data that facilitate a judgement of the extent to which the two different intervention mechanisms were responsible for O. One way of thinking about this is as a second problem of attribution. For internal validity, O must be attributed to T. To facilitate an argument for transferability, O must be attributed to differing extents to the different mechanisms activated by T. Only then can an assessment of the relevant similarity or difference of the causal structures of *φ* and *θ* provide premises for an argument for the extent of the transferability of O from *φ* to *θ*. Therefore, we must give this more precise, theoretically rich interpretation to Subquestion One:

> *What are there systematic differences, if any, between (quasi-)experimental impact evaluation methods regarding the extent to which they report on the barriers and enablers of intervention mechanisms present in the study context and the extent to which they report the degree to which different mechanisms are responsible for changes in outcomes?*

3.3 ANSWERING RESEARCH SUBQUESTION TWO: A CRITICAL REALIST EPISTEMIC COMMUNITIES APPROACH

This section begins by motivating the second research subquestion, defending it as an essential part of answering the overarching research question. The second subsection defends that question against a possible charge that attempting to answer it is over-ambitious. The third subsection sets out the theoretical approach to answering it that I have chosen.

3.3.1 The difficulty of changing practice

As discussed in Subsection 3.2.1, ‘usefulness’ is a key virtue targeted for the account to be generated in response to the primary research question. As such, epistemic rigour is a necessary but not sufficient condition for the success of the account. This is because a strong philosophical argument for a methodological prescription is not sufficient to change research practice. Chapter Two, Section Four illustrated this observation with reference to the failure of ‘there is no gold standard’-type statements to undermine gold standard thinking among development experts.

Other examples can be found across other literatures. For example, consider the complex public health intervention literature. The 2000 Medical Research Council (MRC) guidelines made the strong statement that '[e]valuation of complex interventions *requires* use of qualitative and quantitative evidence' (Campbell et al., 2000, p.1, emphasis added). Nevertheless, Lewin et al.'s (2009, p.1) review of RCTs of complex public health interventions published in English between 2001 and 2003 concluded that qualitative work remained 'uncommon', 'poorly integrated' and 'often had major methodological shortcomings'. Despite the merits of the MRC's guidelines, it is clear that they were not taken up. This episode demonstrates that changing practice requires more than a philosophically compelling argument. A consideration of the insufficiency of a philosophically compelling argument motivates research subquestion two.

3.3.2 Is answering this question too ambitious?

It might be objected that subquestion two is an extremely difficult question to which I am unlikely to be able to respond with much certainty, based on a programme of research that remains within the scope of this PhD. As Chapter Two has outlined, the discourse surrounding contested notions of evidence quality is complex, containing many active agents whose responses to new information will be extremely difficult to model and predict. However, I believe that attempting to give an indicative answer to this question is nonetheless necessary. I could produce work that answers subquestion one and fulfils the conditions for acceptance as a PhD thesis without engaging with that work's ability to change practice. However, such work would not be sufficiently activist to fulfil my responsibilities as a researcher in the *critical* tradition. As Bhaskar says (2008, p.179), following Ravetz (1995), science can have 'social problems'. Sayer (1992, pp.40–43) persuasively insists that criticising these problems is not optional, but stems unavoidably from their identification. By advancing this argument, I position this research in line with the 'critical turn' taken by Sayer and Bhaskar's realism, albeit only to a limited extent (Porter and O'Halloran, 2012, p.18). No endorsement of Bhaskar's later 'transcendental' work is implied. Rather, I follow the critical turn only so far as is necessary to fulfil the responsibility to take seriously and respond to the 'social problem' identified in Chapter Two.

Chapter Two, Section One outlines some of the harms to the development evidence base that result from the appraisal of (quasi-)experimental impact evaluations based solely on the ability of the method used to support a claim to internal validity. Following Sayer, (*ibid*) these harms constitute a social problem whose identification implies its criticism. However, criticism can easily fail to achieve what Sayer (*ibid*) calls its 'emancipatory potential' by falling on deaf ears. Unpersuasive criticism is a common feature of social science discourse. Consider the proclivity of many realists to criticise work as 'positivist' or 'stemming from defective positivist assumptions'. This rhetoric is sometimes employed not only when discussing work amongst persuaded realists, but also when addressing an interdisciplinary audience. It is no wonder that such criticism appears to be unpersuasive of undecided audiences, since as Pawson and Tilley (1997, p.30) themselves put it, 'the term 'positivism' these days has been reduced to a crude term of abuse.' In order to avoid making this sort of error, the research programme described by this thesis seeks to understand the conceptual frameworks and discourses active within the communities whose practices are the subject of critique. The outputs from this research programme can then be framed for and tailored to their specific audiences. This is the 'usefulness' referred to in the primary research question and second sub-research question: In this thesis I use the data generated in answer to the second research subquestion as a lens through which to refract the data generated in answer to the first research subquestion, rendering my conclusions more persuasive to their audience.

3.3.3 Development experts as members of epistemic communities

This research seeks to be useful to the development experts identified in Chapter Two, Section Four. These 'experts' are people who claim intellectual authority over judgements of the quality of (quasi-)experimental impact evaluations of development interventions. An epistemic communities approach is employed because it facilitates the parsing of development experts and their ideas into constituencies whose ideas and motivations can be grouped and understood. Following Haas (1992, p.3), an epistemic community 'may consist of professionals from a variety of disciplines and backgrounds'. However, what unites the community are four sets of shared ideas:

> *'(1) a shared set of normative and principled beliefs, which provide a value-based rationale for the social action of community members; (2) shared causal beliefs, which are derived from their analysis of practices leading or contributing to a central set of problems in their domain and which then serve as the basis for elucidating the multiple linkages between possible policy actions and desired outcomes; (3) shared notions of validity- that is, intersubjective, internally defined criteria for weighing and validating knowledge in the domain of their expertise; and (4) a common policy enterprise-that is, a set of common practices associated with a set of problems to which their professional competence is directed, presumably out of the conviction that human welfare will be enhanced as a consequence' (ibid)*

Of these four sets of ideas, the third is the most important for the purposes of this research. Chapter Two has motivated a claim that there is something deficient about the shared notions of validity in use by epistemic communities that claim expertise over (quasi-)experimental impact evaluations of development interventions. Answering research subquestion one constitutes an attempt to shed some light on this deficiency and provide the building blocks for an improvement of these 'shared notions of validity'. This will take the form of an attempt to give an account of the relative merits of evidence of generated using different methods that considers transferability of findings as well as internal validity. Therefore, to answer subquestion two, it is necessary to identify the different shared notions of validity regarding what counts as a 'high quality' (quasi-)experimental impact evaluation that are currently active in the belief systems of different epistemic communities who study development interventions. Further, it is necessary to investigate the reasons for these shared notions of validity, and any concerns that the community has about them. To borrow the language of Kuhn (1962), as Haas (1992) himself has done, in order to investigate the possibility of contributing to a paradigm shift, it is necessary to identify the 'anomalies' or perhaps even nascent 'crises' that undermine the paradigm. To understand the challenges and limits to altering these shared notions of validity, or paradigms narrowly understood, it is also necessary to understand the roots of their support, not just in scientific puzzles solved, but also in institutional arrangements and the host of other factors that provide incentives for continued adherence to the paradigm.

Haas (1992) was primarily interested in the way in which epistemic communities can influence policy, and not much interested in the ways in which epistemic communities might change over time and interact with advocacy networks, communities of practice, and one another. However, others have explored these dynamics (Keck and Sikkink, 1998; Wenger, 1998). This research will emphasise the dynamic character of the epistemic communities identified and their beliefs, and also understand them as embedded in wider political processes. Following Dunlop (2012, p.8), it will be indispensable not to be naïve about the fact that 'epistemic communities have to be politically proactive players to convey their message, interacting with a multiplicity of other actors where it is to be expected that influence is variable and contingent as wider strategic games are played out'.

In light of the argument of this subsection, we can give a refined, theoretically rich interpretation to subquestion two:

> *For epistemic communities of development experts, what are the shared notions of validity concerning what counts as a 'high quality' (quasi-)experimental impact evaluation? Further, what are the features of these accounts that are valued by members of the community, and what unresolved puzzles or nascent crises undermine them?*

4 Methodology

The subquestions identified in Chapter Three give rise to two distinct but complementary strands of research, requiring different methodological approaches. In this chapter, I examine each subquestion in turn, beginning by outlining the basic research protocol to be employed. I then confront the practical questions that are raised by an attempt to operationalise the subquestion. In answering each of these practical questions I motivate the choice of the research protocols outlined and respond to the methodological challenges that they pose.

4.1 OPERATIONALISING SUBQUESTION ONE

> *What are the systematic differences, if any, between (quasi-)experimental impact evaluation methods regarding the extent to which they report on the barriers and enablers of intervention mechanisms present in the study context and the extent to which they report the degree to which different mechanisms are responsible for changes in outcomes?*

4.1.1 Protocol outline

This subsection outlines the protocol followed during the stage of research Mayoux (2006) refers to as ‘research proper’, after piloting had determined the feasible scope of the activities that could be conducted in that stage. The numbers in parentheses are references to the subsections of this chapter in which arguments are made to support the methodological choices embodied by each part of the protocol.

1) Identify the two intervention-outcome pairs that are best-studied and for which a variety of (quasi-)experimental impact evaluations exist using version 1.1. of the AidGrade dataset.[10] (4.1.4)

[10] The AidGrade v1.1 dataset is downloadable from http://www.aidgrade.org/get-data

2) Augment the two sets of evaluations identified in the AidGrade dataset by searching the 3ie repository of (quasi-)experimental impact evaluations for more evaluations of each of the two intervention-outcome pairs.[11] (4.1.5)

3) For each intervention-outcome pair:

    I. Search the full text of each evaluation and a purposive sample of the wider literature to identify the model or models of intervention causation that predominate and synthesise these to create a realist interpretation of programme theory. (4.1.3, 4.1.6)

    II. Examine this programme theory to identify the contextual factors of relevance, including both properties of the implementation (intervention mechanisms), and properties of the wider context (contextual mechanisms). (4.1.3, 4.1.6)

    III. Code all the evaluations in the set by the extent to which (quantitative) and, where exceptional, the manner in which (qualitative) they report, for the evaluation intervention and population, the contextual factors previously identified. (4.1.3, 4.1.7)

    IV. Compare average scores from (III) between different methods addressing the same intervention-outcome pair, and qualitatively triangulate these results. (4.1.3, 4.1.8)

4) Move iteratively between the methodological literature, the comparison between intervention-outcome pairs, the comparisons within intervention-outcome pairs, and insights from the analysis of individual studies to explain these results. (4.1.8)

5) Combine with the insights from the answer to subquestion two to build an answer to the primary research question. (4.1.8)

4.1.2 Why not a purely analytic approach?

It might be thought that an answer to this subquestion is implied by a precise enough definition of terms. Consider that perhaps, by defining each method precisely we can arrive at a protocol

[11] The 3ie repository of impact evaluations is searchable at http://www.3ieimpact.org/en/evidence/impact-evaluations/impact-evaluation-repository/

that constitutes the method. Then, further, by assessing the extent to which the protocol that constitutes the method generates an understanding of the relationship between outcomes, intervention mechanisms and contextual mechanisms active in the study population, the answer to this subquestion will emerge. This understanding would constitute an answer to subquestion one in that a protocol which generates a detailed description of the relationship between context and outcomes as well as between intervention and outcomes will require much less further knowledge to make an argument for the applicability of findings to a different context. By contrast, a protocol which results in no understanding of the relationship between context and outcomes, and focusses entirely on the relationship between intervention and outcomes will require more further knowledge to make an argument for the applicability of findings to a different context.

In Cartwright's terms, methods that leave the causal structure of the study population φ unexamined will leave much work more work to be done when it comes to making an argument for the applicability of the results to some target population θ. Whereas, methods that require a deep examination and elucidation of the causal structure of φ will provide premises for a relatively straightforward argument about the extent of conformity between φ and θ and therefore about the extent to which intervention T might be expected to cause outcomes O in this new population. I might be expected to examine methods analytically and deduce the extent to which the protocol that constitutes the method requires an exploration of the causal structure of the study population.

This way of approaching the subquestion is rendered implausible by considering the diversity of protocols reported in studies purporting to use the same method. In fact, there is no one-to-one mapping between methods and protocols in practice. When we examine real studies, we find that the protocols deployed might have engaged with, analysed and unpacked the causal structure of the test population to very different extents across studies that all purport to use the same method. This is plainly the case in the difference between 'mainstream' RCTs and so-called 'realist RCTs' proposed by Bonell *et al.* (2012, 2013; Jamal et al., 2015). However, it is also true to various degrees within ‘mainstream’ studies using any method. The protocols employed in the name of a method change over time, and at any given time for any given method there exists a great diversity of protocols being employed by different researchers.[12] This implies that answering subquestion one requires a study of the protocols actually being employed in the name of each method in order to examine the extent to which each method tends in practice to generate useful information about the relationship between context and outcomes that could be used to support an argument for the applicability of findings to some context other than the study population.

4.1.3 Adapting realist programme theory mapping to create a novel method to assess the reporting of contextual information by (quasi-)experimental impact evaluations

The primary methodological challenge in operationalising subquestion one is the fact that there is not a rich methodological tradition that addresses this sort of question. Several methods exist for aggregating information about outcomes from multiple studies. Systematic review, meta-analysis, narrative review and realist synthesis, for example (Littell, Corcoran and Pillai, 2008; Snilstveit, Oliver and Vojtkova, 2012). This research question cannot be answered using a method of aggregating outcomes from multiple studies, however. What is needed is a method of comparing the contextual information generated and described by studies issued from different methods. The ‘useful, systematic account’ targeted by the primary research question, then, can only be provided by a novel method.

Fortunately, realist synthesis provides tools that can be adapted to fill this methodological gap. The method outlined here shares an epistemic approach with realist synthesis but adapts its conceptual framework and methodological toolkit to answer a different type of question. As discussed in Chapter Three, Section Two, for realists, interventions are understood as the embodiment of some theory of the form ‘if we do X in this way, then it will bring about an improved outcome’ (Rycroft-Malone et al., 2012, p.2). This theory is the programme theory. A

[12] See, for example, Angelucci and De Giorgi (2006) and Djebbari and Smith (2008), for two analyses of RCT data about the effects of PROGRESSA to increase household consumption that explore context to very differing degrees.

realist investigation of programme theory will uncover a set of propositions about the effects of intervention mechanisms, moderated by contextual mechanisms, on outcomes. This realist interpretation of the programme theory can be used to build a causal model of how the intervention is supposed to work, moderated by contextual factors, to improve outcomes. In a realist synthesis, after identifying the question and clarifying the purpose of the review, the next stage is to ‘find and articulate the programme theories’ (Pawson et al., 2004, p.vi). This is done by searching for relevant theories in the literature, by drawing up a long list of theories and then by grouping, categorising and, where possible, synthesising those theories. This process of mapping the programme theory is used to create an ‘evaluative framework’ to guide the rest of the review. Interactions between factors in the review framework are identified to be populated with evidence to build an answer to the question ‘what works for whom, in what circumstances, in what respects, and how?’ (*ibid*, p.v).

The process of mapping the programme theory present in the literature to create the ‘evaluative framework’ is a methodological tool that can be adapted from realist synthesis to help answer subquestion one. In a realist synthesis, the realist causal model of the evaluative framework, built through an analysis of programme theory, is expressed as a set of context-mechanism-outcome (CMO) configurations.[13] Answering research subquestion one requires comparing (quasi-)experimental impact evaluation methods regarding the ‘extent to which they report on the barriers and enablers of intervention mechanisms present in the study context and the extent to which they report the degree to which different mechanisms are responsible for changes in outcomes’. Both of these can be derived from the CMO configurations created through programme theory mapping. Because CMO configurations explain how intervention mechanisms combine with context to produce changes in outcomes, the contextual barriers and enablers of intervention mechanisms are implied by them. Therefore, for a set of evaluations of an intervention-outcome pair, realist programme theory mapping can be used to aggregate and

[13] Authors in the literature also sometimes disaggregate ‘context’ or ‘mechanism’ to produce four-element configurations such as context-mechanism(resource)-mechanism(reasoning)-outcome configurations (e.g. Dalkin et al., 2015), or context-intervention-mechanism-outcome configurations (e.g. Denyer, Tranfield and van Aken, 2008).

represent the theory or theories of intervention causation that underpin evaluations in the set, and this representation of theory can be interpreted to provide a list of contextual 'barriers and enablers of intervention mechanisms.' This list is one of two elements that are required in order to assess (quasi-)experimental impact evaluations issued from different methods in order answer research subquestion one. The second element required is a way of determining the extent to which evaluations have reported the degree to which different mechanisms are responsible for changes in outcomes. This assessment requires a description of the information that must be reported in order to determine the extent to which different intervention mechanisms have been activated in a given context. This information can also be derived from the programme theory represented as CMO configurations. This might be because the CMO configurations specify intermediate outcomes that are affected by some mechanisms but not others, or because different mechanisms will change final outcomes for different subgroups of intervention recipients. Chapters Five and Six demonstrate how this information was derived from the programme theory map for each of the two cases of intervention-outcome pairing studied in this research project. The information required to attribute changes in outcome proportionally between mechanisms can be combined with the information required to specify the existence of barriers and enablers of those mechanisms in a given context. This creates a list of contextual markers that are sufficient premises, if reported, for an argument for the extent of the transferability of treatment effects to some target context.

To illustrate how realist programme theory mapping can generate the information required for an attempt to answer research subquestion one, consider the case of conditional cash transfers (CCTs) targeting increased school enrolment for children in recipient households. A realist programme theory mapping exercise could be conducted to search for, list, group and synthesise the theories that underpin evaluations of CCTs for school enrolment. This would uncover that an essential part of programme theory is that investments in children's education are constrained by households' available financial resources (Fiszbein and Schady, 2009). It would also uncover the theory that one mechanism through which CCTs boost enrolment is the changes in household decision making that result from a transfer of resources to households. Low levels of available household financial resources would therefore have been identified as a contextual enabler of one mechanism through which CCTs operate to increase school enrolment. A list of all of the major contextual barriers and enablers for each mechanism could be drawn up in this way. These barriers and enablers could be combined with the information required to discern which mechanisms were active to which extent. In this case, that would be a disaggregation of treatment effects over household available resources amongst other things.[14] This would create a full list of the markers of intervention causation in context (MICCs) that should be reported by an evaluation in order to facilitate an argument for the transferability of results to some other setting.

An advantage of adapting the tool of programme theory mapping from realist synthesis is that several sets of guidelines exist to help researchers in using this tool. Since Pawson and Tilley's (1997) original work describing 'realistic evaluation' and the introduction of realist synthesis in an ESRC methods paper (Pawson et al., 2004), more concrete guidelines for realist synthesis have been produced. The most recent and robust of these is the set of quality standards for realist synthesis issued by the RAMESES project (2014). I will be guided in my construction of causal models for each intervention-outcome pair selected by these guidelines. However, there is one aspect of these guidelines that I will diverge from. The guidelines ask researchers to ensure that '[t]he final realist programme theory comprises multiple context-mechanism-outcome configurations [CMOs] (describing the ways different mechanisms fire in different contexts to generate different outcomes) and an explanation of the pattern of CMOs' (RAMESES, 2014, p.4). This recommendation, and some others, seem to imply the creation of a singular, unified programme theory, different aspects of which are to be tested by subsequent stages of a realist review. In fact, as Pawson *et al.* articulate in the original ESRC methods paper, 'the ideas unearthed in a theory mapping exercise will be many and varied' (p.35) and will not necessarily constitute one unified programme theory. The list of contextual factors of relevance to an argument for application of a study's findings to a different context will be

[14] See Chapter Five for a full treatment.

contested in the literature. My theory-mapping exercise will therefore uncover a list that contains factors with different levels of support in the literature. Before assessing the extent to which studies in the literature generate and report these factors for their interventions and contexts, I will reduce this list to those factors that are explicitly part of the causal models that are most popular in the literature, or are implied by a realist interpretation of those same theories. Including some causal factors and excluding others introduces a potential source of bias in my analysis by giving me one more 'researcher degree of freedom' (Simmons, Nelson and Simonsohn, 2011). It will therefore be essential, when mapping the literature, to be systematic and transparent about the ways in which I judge whether an element of programme theory has widespread or narrow support.

So, using realist programme theory mapping can provide a method for building a set of the markers of intervention causation in context (MICCs) that should be generated and reported by evaluations of a given intervention-outcome pair. The extent to which this information is generated and reported by evaluations can then be compared between different methods. As well as comparing results between sets of evaluations using different methods to study the same intervention-outcome pair, I can study multiple intervention-outcome pairs and compare results between sets of evaluations for different intervention-outcome pairs. This is important in order to identify and examine possible sources of bias resulting from the choice of intervention-outcome pair. This is described in more detail in Subsection 4.1.4.

A consequence of developing a novel method in the operationalising of research subquestion one is that the subquestion becomes two-fold. On the one hand, it must be established that the novel method is successful. Then, what it tells us about the systematic differences between methods can be interrogated. This implies an operationalised interpretation of research subquestion one:

1.

a) Can realist programme theory mapping be adapted to create a tool to assess the transferability of (quasi-)experimental development impact evaluation results?

b) If so, what can it tell us about the systematic differences, if any, between (quasi-)experimental impact evaluation methods regarding the extent to which they report on the barriers and enablers of intervention mechanisms present in the study context and the extent to which they report the degree to which different mechanisms are responsible for changes in outcomes…

i. as they are currently used?

and

ii. as they might be used?

part b) of this question is split into two subcomponents to reflect the observation of Subsection 4.1.2 that the protocols employed in the names of methods are not fixed. Therefore, as well as assessing (quasi-)experimental development impact evaluation methods as they are currently used, a useful, systematic account of the relative merits of these methods will be able to offer suggestions for how those methods could be used better.

4.1.4 Using the AidGrade database to identify two intervention-outcome pairs to study

As the previous subsection has described, comparing the extent to which contextual information is provided by studies across a set of methods requires restricting the scope of the analysis to groups of studies dealing with the same intervention-outcome pair. However, selecting one or more intervention-outcome pairs before conducting a comparison of methods raises a methodological challenge. There is a possibility of bias in the results that arises from limiting the analysis to one or more intervention-outcome pairs. On the other hand, examining the total universe of evaluations of the effectiveness of development interventions is not possible, nor is randomly selecting studies, if those studies must describe the same intervention-outcome pairs in order to be comparable. Therefore, it is necessary to select only a limited number of intervention-outcome pairs to study, but desirable to select intervention-outcome pairs that are contrasting in order to minimise foreseeable sources of bias.

The quantitative comparison of MICCs reported will be qualitatively triangulated and deepened as described in Subsection 4.1.7. However, the quantitative element will still have to be backed by a number of studies in each group that provides sufficient statistical power for a persuasive comparison of average scores across groups. This number cannot be precisely determined in advance, because it depends on the variation in scores both within and between groups (Cohen, 1992). When selecting intervention-outcome pairs, then, it is desirable to select those pairs that are best-studied in order to compare within the pairing over as many evaluations as possible.

Pairing of intervention and outcome is not normally a dimension on which studies are catalogued. For example, the 3ie repository aims to be an exhaustive collection of (quasi-)experimental impact evaluations of development interventions and currently contains 4,635 studies. However, while studies are categorised according to many characteristics including the method employed and the sort of intervention evaluated, outcome variables are not recorded and reported for every study. Therefore, no categorisation by intervention-outcome pair is possible. In order to identify suitable cases of intervention-outcome pairs that have been studied extensively using a variety of different (quasi-)experimental impact evaluation methods, a database of (quasi-)experimental impact evaluations that does contain information on intervention type and outcome is the ideal tool. Fortunately, the AidGrade repository of (quasi-)experimental impact evaluations provides a method of identifying such cases.

Constructed between 2012 and 2014, the most recent version (1.3) of the AidGrade database of (quasi-)experimental impact evaluations contains 635 studies identified during the production of 20 meta-analyses or systematic reviews.[15] The most recent studies included in the AidGrade dataset were published in 2013. Version 1.3 of the dataset is not available publicly. However, version 1.1 can be downloaded from AidGrade's website. This version contains all of the same

[15] All projects were begun with the intention of producing meta-analyses, but some 'did not have enough comparable outcomes for meta-analysis and became systematic reviews' (Vivalt, 2015, p.467).

evaluations as version 1.3, but not all of the same variables for those evaluations. For the purposes of this research project, that makes the two versions equivalent. The process of intervention selection employed by AidGrade involved the proposition of interventions of interest by researchers and the voting on interventions by members of the public. This results in a list of interventions that are considered to be highly relevant by both researchers and the public. Within these interventions, the process of intervention selection prioritised the most-studied. By selecting the most-studied intervention-outcome pairs within the AidGrade database, I can select cases of intervention-outcome pairs that are highly relevant and well-enough studied to allow for comparisons between methods in their treatment of contextual factors.

A necessary property of the AidGrade dataset for my purposes is that it is permissive in terms of the (quasi-)experimental impact evaluation methods selected, including all forms of quasi-experimental and experimental methods. Matching studies, difference-in-differences studies, instrumental variables approaches, regression discontinuity designs and randomised controlled trials are all represented (Vivalt, 2015). Therefore, the AidGrade dataset can be used to identify all of the (quasi-)experimental impact evaluations conducted up to 2013 for a large number of relevant and well-studied intervention-outcome pairs. Sorting the AidGrade dataset by intervention-outcome pair and ranking by number of studies reveals that the top two most-studied intervention-outcome pairs in the dataset are CCTs for school enrolment and deworming for weight with 29 and 18 evaluations in the dataset respectively.

As discussed earlier in this subsection, it is desirable to study as many intervention-outcome pairs as possible. However, early piloting determined that the method was extremely time-consuming. Each evaluation in each set of evaluations of the same intervention-outcome pair had to be carefully examined in its entirety twice; once to extract relevant theory and then a second time to score against the list of MICCs described. In addition, for each set, full text examination of a large number of theoretical references was necessary.[16] It is therefore a

[16] 25 references for the first set, 39 for the second.

regrettable but necessary limitation of this research project that only two pairings of intervention and outcome could be examined. These two cases of sets of evaluations must be treated as case studies, rather than as data points in a statistical analysis in which $n = 2$. Within cases, there are enough evaluations from diverse methods to make the quantitative analysis of the number of MICCs reported by each evaluation informative, in addition to the qualitative analysis of differences between methods. Between cases, a qualitative analysis is the only form of analysis supported by the low number of cases.

The risk of bias resulting from studying only two cases of an intervention-outcome pairing is reduced by the fact that the two cases studied are contrasting cases. We might expect one very important source of bias in an analysis of (quasi-)experimental impact evaluation methodology to be the broad class of interventions to which the specific intervention studied belongs. On this dimension, a case built on the analysis of a set of deworming interventions contrasts with a case built on the analysis of a set of CCT interventions. The former interventions belong to the broad class of public health interventions whereas the latter belongs both to the broad class of education interventions and to the class of social protection interventions. Not only do these three classes of intervention contrast with each other, they are also the three most studied classes of intervention. Between them, they account for 65% of all the (quasi-)experimental impact evaluations in the 3ie repository of (quasi-)experimental impact evaluations, which aims to be an exhaustive collection (Sabet and Brown, 2018). Another important source of bias in the study of (quasi-)experimental impact evaluation methodologies might be the disciplinary backgrounds of the architects of the evaluations. Sabet and Brown report that 47% of all the (quasi-)experimental development impact evaluations in the 3ie repository were published in health journals. 53% were published in social science journals, as a working paper, or as a report. As working papers and reports are more commonly published by social scientists, Sabet and Brown interpret this as a roughly 50-50 split in authors of (quasi-)experimental development impact evaluations between social scientists such as economists and those from a public health or epidemiological background. It therefore reduces the expected bias of this analysis that the two cases studied are dominated, respectively by economists on the one hand and public health researchers on the other.

4.1.5 Extending the sample for each case using the 3ie (quasi-)experimental impact evaluation repository

A major limitation of the AidGrade dataset, as has been mentioned, is that it only contains evaluations published in 2013 or earlier. It might be argued that an assessment of the relative strength of (quasi-)experimental development impact evaluation methods should include the most recent evaluations possible. Impact evaluation practice is constantly evolving and improving, and so a 2021 assessment of methods based on data up to 2013 might be considered out of date at the time of publication. In order to address this concern, the sets of evaluations identified using the AidGrade dataset are extended with more recent evaluations identified using the 3ie repository of (quasi-)experimental impact evaluations. This repository was created using a systematic search strategy, snowball reference following and crowdsourced additional contributions to attempt to exhaustively catalogue studies of development interventions employing an 'impact evaluation' methodology (Cameron, Mishra and Brown, 2016). By 'impact evaluation' methodology is meant, as in the AidGrade dataset, an attempt to measure the impact of an intervention on outcomes using either an experimental or quasi-experimental design to establish a reference group of potential beneficiaries who did not receive the intervention and are minimally systematically different from the group of beneficiaries who did receive the intervention (White, 2010). The repository was updated in 2018 and currently contains 4,635 studies. The main methods employed are the RCT (1,985 studies), difference-in-difference design (505 studies), propensity score matching (497 studies), instrumental variable design (239 studies), and regression discontinuity design (91 studies). As well as the method employed, studies are coded by the sort of intervention which they evaluate. For example, there are 255 studies assessing conditional cash transfers.

It is unfortunate that 3ie do not make the full dataset of evaluations available to researchers. However, it is possible to query the data using the web-based search at http://www.3ieimpact.org/en/evidence/impact-evaluations/impact-evaluation-repository/. Using this search function, it is possible to filter evaluations by 'Sector'. Fortunately, one of the sector tags employed by 3ie is 'Conditional Cash Transfer'. However, there is no sector tag that corresponds to deworming interventions. Therefore, the process of updating the two sets of evaluations with more recent evaluations from the 3ie repository was slightly different. For conditional cash transfers for school enrolment, all of the evaluations tagged with the sector 'CCT' were identified. The resulting list of 256 short entries was screened manually for date of publication, identifying 62 entries listing an evaluation with a publication date of 2013 or later. Abstracts of studies published up to 2013 were screened to identify any evaluations not included in the AidGrade database but which might be relevant. This identified one study which was selected for inclusion. For the 62 evaluations published in 2013 or later, full texts were examined to ascertain whether school enrolment was a reported outcome variable. To identify further evaluations of deworming for child weight, it was necessary to use the text search function to identify database short entries related to deworming evaluations. In order to achieve this, a series of potential search terms were identified and piloted, and any term that increased the number of returned results was included in the final search string. 47 candidate evaluations were identified in this way. As with the first case, the short entries describing these evaluations were then screened to examine abstracts of studies published in 2012 or earlier to check for evaluations not included in the AidGrade database. In this case, no new evaluations were identified. The full texts of all evaluations published in 2013 or later were examined to include only evaluations of deworming interventions that reported child weight or an equivalent outcome variable. More detail on this process for each of the cases is available in Chapters Five and Six.

4.1.6 The process of programme theory mapping

Having generated a set of all the evaluations contained in the AidGrade and 3ie databases for each intervention-outcome pair, the next stage in the method was to identify the accounts of intervention causation that underpin evaluations in each set. As Subsection 4.1.3 has described, the process of programme theory mapping can be adapted from realist synthesis in order to achieve this objective. This is a retroductive process of moving between evaluations in the set, the wider theoretical literature and the programme theory map under construction until that theory map adequately describes the theory or theories that underpin evaluations in the set. This subsection describes the practical steps taken to operationalise this process.

The purpose of uncovering the theory or theories underpinning evaluations in each set is to use that theory or those theories to build a list of the markers of intervention causation in context (MICCs) implied by it. This list is then used, as described in Subsection 4.1.3, to assess each evaluation on the extent to which it reports the MICCs implied by the theory underpinning that evaluation. Evaluations that share a theory of intervention causation can be compared with each other more readily than evaluations that do not share a theory of intervention causation. Therefore it is desirable to aggregate and synthesise the theories of intervention causation underpinning evaluations in the set where this is possible. If two or more competing schools of thought are discovered, the evaluations must be separated into two or more subsets of evaluations from each school of thought. Each evaluation can then be compared with other members of the subset to which it belongs. If any evaluation were found to be based on a theoretical framework not shared by other evaluations in the set, it would have to be excluded from the quantitative comparison of marker reporting and would only be included in the qualitative analysis component. It was therefore necessary to investigate the full text of every evaluation in order to accurately group evaluations and to uncover any evaluations based on non-standard theories.

In practice, this meant that for each evaluation in each set, the full text was examined and programme theory extracted. This theoretical information was added to a working spreadsheet and reconciled with material already present in the spreadsheet where possible. At the end of this process, for each superset of evaluations belonging to a case, a spreadsheet containing elements of programme theory had been created. As initial piloting of the theory extraction process with a subset of the evaluations had revealed, evaluations in both sets contained varying levels of theorising about mechanisms. Some evaluations, especially those designed to test an aspect of programme theory, contained highly developed programme theory sections. However, other evaluation documents contained very little elaboration of theory, mentioning theory very briefly and referring the reader to (presumably) more theoretical works. As some of these more theoretical works were not in the set of evaluations, it was clear that a programme theory map based only on the information contained in evaluations in the sets would not be sufficient, therefore it was necessary to consult the wider literature for each case.

Both the conditional cash transfers literature and the deworming literatures are extensive and consulting every document containing some information about intervention theory was not feasible. Moreover, it would not have been desirable for the purposes of this research project, the goal of the theory mapping process being restricted to identifying the programme theories *that underpin evaluations in the set*. Therefore, as well as a close reading of every paper in the evaluations set, a snowballing strategy was adopted to explore the trees of theoretical references stemming from root papers within the evaluations set. In addition, to increase the efficiency of this search, existing literature reviews referencing evaluations in the set and systematic reviews containing evaluations in the set were identified. Initial piloting revealed these to be rich sources of theoretical information that could increase the efficiency of the search process. That resulted in a three-stage literature review strategy. For each case, first, full texts of all evaluations in the set were examined, then systematic or narrative reviews identified through web searches. Finally, the theoretical references of reviews were investigated using a snowball sampling design that is described in the following paragraph.

It was not possible to examine the full text of every theoretical reference of every evaluation in the set, given the capacity available. Therefore, a snowball sampling procedure was employed that began by selecting an evaluation in the set at random.[17] References relating to programme theory from this evaluation, the ‘root evaluation’, were collected to create a set of further theoretical references. The referenced documents were then read closely and all information

[17] Random selection was achieved by generating a new variable in the dataset of studies populated with an independent random value between 0 and 1 for each study. Studies were then sorted on this variable and selected in order.

relating to programme theory was extracted from them and used to update the theory map in progress. For any major points of theory that were referenced to additional documents, these were also added to the set of theoretical references to be followed. This extension could continue more than once, until the theoretical allusion of the root paper had been sufficiently elaborated to be considered well-grounded in the literature. Once this iterative branching search strategy had been exhausted, a different root evaluation was selected at random, and any new theoretical references followed and investigated. This retroductive movement between programme theory map and source documents was continued until no significant additions had been made to the programme theory map for three consecutive root evaluations. This method is inspired by the well-established practice of continuing a qualitative investigation up until a point of ‘saturation’ or ‘data adequacy’ (Morse, 1995; Glaser and Strauss, 2017). In realist terms, the point at which the investigation of a new phenomenon consistently fails to present a challenge to the working description of the set of phenomena to which it belongs can be understood as the point at which that description has achieved ‘practical adequacy’ (Sayer, 1992).

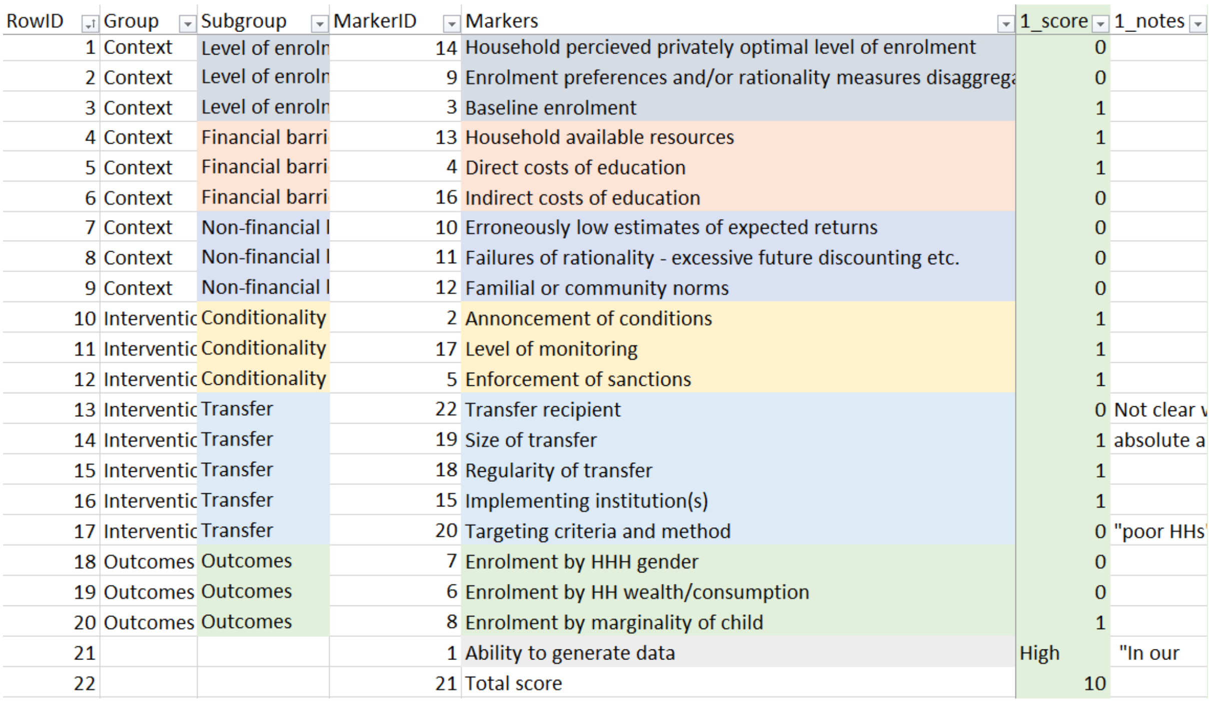
For each case, the programme theory mapping exercise described above resulted in a spreadsheet containing programme theory elements grouped by overriding programme theory or ‘school of thought’. In practice, for both cases the set of evaluations were underpinned by a remarkably homogeneous theoretical literature. Therefore, it was possible to synthesise the programme theories underpinning evaluations in each set to create a single account of programme theory relevant to all evaluations in each set. Chapters Five and Six explain this process in more detail for each case. These theories were interpreted through a realist lens to be represented as context-mechanism-outcome (CMO) configurations as described in Subsection 4.1.3. From these CMO configurations, a list of markers of intervention causation in context (MICCs) for each case was derived. By construction, the elements of this list provide adequate premises for an argument for the transferability of the results of an evaluation underpinned by the theory from which the markers had been derived.

4.1.7 Coding evaluations

As the previous subsections have outlined, for two cases defined by two intervention-outcome pairs, I surveyed the literature to create a list of MICCs whose reporting is necessary for an argument for the transferability of the findings of an evaluation reporting results for that intervention-outcome pair. The next stage of my analysis was to assess the evaluations within each set to assess the extent to which those evaluations generated and reported the necessary contextual markers for their interventions and contexts. The literature on study aggregation methods provides models for how this process can be completed. For example, the AidGrade Coding Manual (2013a) and Meta-Analysis Process (2013b) provide very great detail on how AidGrade reviewers coded outcome variables from studies for inclusion in meta-analyses. Borrowing from this literature, I created two Microsoft Excel spreadsheets for coding, one for each case. Each spreadsheet was designed with one row for each MICC plus an extra row for 'Ability to generate data' and one for 'Total score.' Leftmost columns recorded various information about each marker including group, subgroup, the name of the marker and a 'Marker ID' of ascending integers 1 to *N* where *N* = the number of markers + 2 for 'Ability to generate data' and 'Total score.' Right-hand columns were labelled 1_score, 1_notes, 2_score, 2_notes, … *N*_score, *N*_notes where *N* = the number of studies for the case. Cells in '*n*_score' columns were formatted to accept only entries of 1 or 0. Cells in '*n*_notes' columns were formatted to accept strings of text to allow me to qualitatively elaborate on the judgements entered in the corresponding *n*_score cell.

In summary, I developed a spreadsheet with one row for each marker, plus one for 'Ability to generate data' and one for 'Total score.' In addition to columns for describing each marker there were two columns for each evaluation in the set, a quantitative score column and a qualitative detail column. The resultant spreadsheet was formatted as shown in Figure 4.1, with colour-coding used to make the sheet clearer at a glance to try and reduce errors in coding. This spreadsheet structure facilitated coding and allowed for a level of quantitative analysis and comparison between studies, deepened though qualitative triangulation of these results (Yeung, 1997; Mayoux, 2006).

Figure 4.1: Coding spreadsheet formatting



RowID	Group	Subgroup	MarkerID	Markers	1_score	1_notes
1	Context	Level of enroln	14	Household percieved privately optimal level of enrolment	0	
2	Context	Level of enroln	9	Enrolment preferences and/or rationality measures disaggrega	0	
3	Context	Level of enroln	3	Baseline enrolment	1	
4	Context	Financial barri	13	Household available resources	1	
5	Context	Financial barri	4	Direct costs of education	1	
6	Context	Financial barri	16	Indirect costs of education	0	
7	Context	Non-financial I	10	Erroneously low estimates of expected returns	0	
8	Context	Non-financial I	11	Failures of rationality - excessive future discounting etc.	0	
9	Context	Non-financial I	12	Familial or community norms	0	
10	Interventic	Conditionality	2	Annoncement of conditions	1	
11	Interventic	Conditionality	17	Level of monitoring	1	
12	Interventic	Conditionality	5	Enforcement of sanctions	1	
13	Interventic	Transfer	22	Transfer recipient	0	Not clear v
14	Interventic	Transfer	19	Size of transfer	1	absolute a
15	Interventic	Transfer	18	Regularity of transfer	1	
16	Interventic	Transfer	15	Implementing institution(s)	1	
17	Interventic	Transfer	20	Targeting criteria and method	0	"poor HHs'
18	Outcomes	Outcomes	7	Enrolment by HHH gender	0	
19	Outcomes	Outcomes	6	Enrolment by HH wealth/consumption	0	
20	Outcomes	Outcomes	8	Enrolment by marginality of child	1	
21			1	Ability to generate data	High	"In our
22			21	Total score	10	

The process of identifying evaluations in the AidGrade data and then extending that set of evaluations by searching the 3ie repository created a dataset of studies identified for each case containing variables populated with study characteristics such as title, publication year, authors etc. For this dataset an ID field was generated, populated with a random number, sorted, and then re-coded 1, 2, 3 … *N* in the order of sort. This variable was then used to link studies in each set with the columns recording observations in the coding spreadsheet, with the number used to determine the order in which studies would be coded for the reporting of MICCs. This made the order of coding random for studies in each set. This was done to ensure that any differences in coding associated with order of coding, as well as being minimised by careful attention and double-coding, would be randomly distributed between evaluations in the set.

The quantitative comparison of the reporting of contextual information between groups of studies could potentially have made use of a weighting of contextual factors by importance. Such a weighting would only have been employed if it were justified by the conceptual model(s) identified during the programme theory mapping and synthesis process. A conceptual model of intervention causation which met Cartwright's (2007) specification for adequacy would include a 'rule of combination' that expresses, even if only approximately and probabilistically, what effect on outcomes causal mechanisms and moderating factors should have. Uncovering such rules of combination would allow us to determine the relative importance of conceptual factors in terms of the relative magnitude of their effects on outcomes. In practice, the programme theories that underpinned evaluations in both sets were not sophisticated enough to include, even implicitly, such rules of combination. In the absence of a justification for weightings derived from rules of combination, contextual factors could only be equally weighted in my analysis.

Double-coding by different researchers, followed by a process of reconciliation, is considered an important element of coding practice in order to avoid errors and to reduce biases resulting from one reviewer's way of interpreting studies. As I was not able to employ a second a researcher to double-code studies, this practice was not be available to me. However, I was still able to reduce errors and omissions by double-coding myself and then reconciling the two sets of answers. To reduce the time spent on this process, and therefore increase the number of studies it was possible to include, I restricted this double-coding to the binary '*n*_score' fields for each study. In case of conflict between the two sets of coded variables, I also revisited the string fields recorded in '*n*_notes'. Otherwise, I did not re-enter my qualitative assessments of the reporting of contextual factors. The two spreadsheets of MICC reporting were exported as csv files once coding was complete, and imported into STATA for data analysis using the .do files included in this thesis' accompanying data files. These are available at https://mattjudendotcom.files.wordpress.com/2021/04/mjuden_thesis_data_and_code.zip with data uploaded to the UK Data Service where it has been deposited in accordance with the conditions of my ESRC studentship under Project ID 204604.

4.2 OPERATIONALISING SUBQUESTION TWO

> *For epistemic communities of development experts, what are the shared notions of validity concerning what counts as a 'high quality' (quasi-)experimental impact evaluation? Further, what are the features of these accounts that are valued by members of the community, and what unresolved puzzles or nascent crises undermine them?*

4.2.1 Protocol outline

This subsection outlines the protocol followed in order to generate an answer to subquestion two. The objectives, A to C, are not sequential tasks. Rather, they are goals to be pursued during an iterative movement between the data-generating strategies I-III. This data-generation was intended to continue until a point of practical adequacy (Sayer, 1992).[18] In the following subsections arguments are made to support the methodological choices embodied by each part of the protocol.

Objectives

A. Identify epistemic communities that claim authority over judgements of the quality of development intervention evaluation evidence.
B. Identify their 'shared notions of validity' concerning what counts as a 'high quality' (quasi-)experimental impact evaluation.
C. Identify the features of these accounts that are valued by members of the community as well as any unresolved puzzles or nascent crises that put pressure on them.

Data-generation strategies

I. Consultation of the literature, both academic and otherwise published by authors claiming authority over what counts as a high-quality evaluation of a development intervention.
II. Semi-structured interviews with experts on what counts as a high-quality evaluation of a development intervention.

4.2.2 What are the most time-efficient methods of answering subquestion two?

When choosing between methods, Sayer (1992, p.4) urges us to 'imagine a triangle whose corners are method, object and purpose' and remember that 'each corner needs to be considered in relation to the other two'. My purpose for this part of this programme of research is to answer the operationalised interpretation of subquestion two. That is to say, it is to *identify* and *describe*

[18] See subsection 4.1.6 for more detail on this concept and Chapter Eight, Section One for the barriers to achieving it in practice.

A) a set of epistemic communities, B) the shared notions of validity which serve partly to define those communities and C) the features of these accounts that are valued by members of the community, as well as the unresolved puzzles or nascent crises undermine them. Having identified the objects and purpose of this part of my research helps to limit the set of methods that are appropriate. This set of potential methods can be further reduced by considering the available sources relevant to this research.

Members of the epistemic communities implicated in the study of development interventions publish a lot of material. That part of their published material that deals with questions of evidence quality takes the form both of academic, peer-reviewed work, and so-called grey literature.[19] Both will be of relevance to an answer to subquestion two. Therefore, an iterative engagement with the literature is an essential component of this research.

The interpretation of subquestion two given above also implies an engagement with community members' belief systems. These may not be adequately expressed in community members' writings, requiring a level of access to community members beyond their writings (Haas, 1992). The approach adopted was to conduct semi-structured interviews with experts on the evaluation of development interventions. Selection of participants is covered in the next subsection. Semi-structured interviews were the most appropriate form for these interviews for two reasons. Firstly, these interviews were intended to generate data of a very specific type in a few areas. Rather than being unstructured projects of discovery, the data generated are heavily theoretically informed as described in Chapter Three, Subsection 3.3. Secondly, a highly structured questionnaire would not have been appropriate, as some freedom was required to 'probe and expand the interviewee's responses' in order to 'achieve depth' (Rubin and Rubin, 2005, p.88). The semi-structured interview is the format that allows me to impose some structure by prompting discussion of the areas that I am interested in, while allowing flexibility for the elaboration and exploration of complex subjects.

[19] See Befani (2016) for an example of the sorts of practitioner-facing non-peer-reviewed methodology guides that are very common.

The process of interviewing is fraught with well-documented methodological challenges (Weiss, 1995, chap.4; Taylor and Bogdan, 1998, chap.4). These can partly be addressed by reassuring participants in the research of the anonymity of their contributions, and clearly explaining my motives and intentions (*ibid*). For the purposes of answering subquestion two, my motives are essentially to understand participants' ways of thinking in order to present my answer to subquestion one to them in the way that will be most comprehensible to them, and is most likely to chime with the existing knowledge puzzles in their way of viewing evidence quality. My motives are therefore essentially non-threatening; I seek to be useful to participants, and invite them to help me in that goal. It is for this reason that I was confident in overcoming the challenges to productive interviewing, and that I refer to the subjects of interviews as 'participants' throughout this section. I also use the terminology of 'participant' to reflect the fact that preliminary interviews with participants were conducted during the scoping stage of this research. These interviews helped to shape this thesis through the suggestions of participants, who recommended reading materials and advised on the feasibility of the methodological approaches suggested to them. I intend to continue this engagement beyond the 'writing up' stage of this project, in particular by using participatory identification of gaps, limitations and dissemination techniques during the dissemination stage.

4.2.3 Should participant observation have been employed?

Dunlop (2012) considers some of the beliefs of community members relevant to the delineation and exploration of epistemic communities to be 'sensitive information' and, following Haas, recommends the 'soaking and poking' approach to data collection, originally formulated by Fenno (1986). This approach could be taken to imply a level of engagement with research subjects that borders on participant observation. Fenno cultivated deep relationships with research subjects, who were observed as they engaged in activities that had not been pre-determined to be relevant to research. Indeed, Fenno (*ibid*, p.3) believed that it was essential to his observation of the behaviour of US Senators to '[watch] them in two contexts – at home and in the capital city.'

It might be objected that the research protocol outlined in Subsection 4.2.1 is deficient in not including participant observation within the epistemic communities studied. Historically, participant observation has been considered a sort of ‘gold standard’ of its own (Atkinson and Coffey, 2003). Becker and Greer (1958, p.133) famously even went so far as to claim that ‘the most complete form of the sociological datum … is the form in which the participant observer gathers it.’ However, Becker and Greer (1958) themselves clarified that participant observation was only a superior method of data collection for a very specific kind of research, dealing with ‘specific and limited events’. Returning to Sayer’s (1992) insistence on the primary importance of research purpose to method selection, I reflect that a survey of the Anglophone epistemic communities that study the effectiveness of development policy interventions is a research purpose that requires substantial breadth. In the time available, this is incompatible with a method of the narrowness and depth that participant observation implies. It is for this reason that research employing the epistemic communities approach has generally limited its engagement with epistemic community members to literature review, archival research and interviews (Dunlop, 2012). I have followed in that tradition.

4.2.4 Sampling and interview process

This subsection describes the sampling and interview processes in accordance with the Consolidated Criteria for Reporting Qualitative Research (COREQ) developed by Tong et al. (2007). All of the checklist elements required for COREQ are reported in this subsection, with the exception of the theoretical framework, which is reported in the previous chapter.

Random sampling of interview subjects was neither possible nor desirable. It was not possible because no sample frame existed, nor could one be constructed in advance, that contained all of the potential members of epistemic communities of relevance. It was not desirable because the purpose of this aspect of my research was not to describe the total population of experts in proportional terms. I did not need to discover what proportion of experts believed X. Rather, my purpose was to identify the major epistemic communities, their shared notions of validity, and the strengths and weaknesses that community members consider these notions of validity to have. This purpose required selecting participants on the basis of their difference. Participants must still be identified in order to be selected, however. In the absence of a sample frame, this poses a methodological challenge. The standard response is to employ ‘snowball sampling’, in which new participants are identified by existing participants as a part of the researchers’ engagement with them (Biernacki and Waldorf, 1981). The major weakness of this approach is that the selection of new participants is heavily biased, with new participants likely to be similar to existing participants. This difficulty can be mitigated by seeking out new participants that are different in respects that existing working theory suggests might be important, and beginning new chains of referral from these new participants (*ibid*). This is the approach that I took. My chains of referral began with existing connections who are members of the epistemic communities of relevance to my research, and were diversified later, after every iteration of retroductively moving from data generation to interpretation and back.

In total, 12 interviews were conducted with experts on (quasi-)experimental development impact evaluation. Two were known to me to a low degree in a purely professional context. The others were totally unknown to me at the outset of this research project. The interviewees were selected using a mixture of snowball sampling and sampling on difference. It would have been optimal to continue sampling until a point of theoretical saturation, where new interviews were consistently failing to generate new insights (Taylor and Bogdan, 1998). However, the wealth of information generated by 30 to 45 minute semi-structured interviews rendered this unrealistic. Every new interview generated a lot of new information. Time constraints therefore made it impossible to conduct enough interviews to arrive at a point of theoretical saturation.

Interviewees were approached by email, either on the basis of a recommendation from another participant or as a result of identification through web searches to identify staff at suitable organisations. Rate of non-response to the first email was moderately high with a success rate of 15/62 emails. Of these 15 first responses, 12 resulted in interviews. Participants were sent some basic information about myself, the research, what topics would be covered in questions, and how the data generated would be used. This is available in Appendix C. If they agreed to participate, participants were sent a consent form, an example of which is available in Appendix D.

Interviews were conducted at participants’ places of work either face-to-face in an empty meeting room, or in one case in a co-working space’s communal café area, or via Skype. No other persons were present in any of the interviews. Interviews were conducted in a semi-structured fashion, using a very basic guide to ensure that seven key topic areas were covered. This guide is available in Appendix E. No repeat interviews were carried out. All interviews were audio recorded. Some notes were taken, though these were not found to be useful and were discarded in favour of analysis of transcripts, which were not returned to participants for comment or correction. Participants were offered this option, but did not desire to take it up. It is likely that participants’ lack of desire to correct or comment upon their accounts is related to the fact that participants and their organisations were promised anonymity.

4.2.5 How was data interpreted?

Interpreting the data was not easy. Even the identification of epistemic communities has been found challenging by many researchers. For example, Wright (1997, p.11) cautions that ‘actually identifying these communities … can be a difficult process’. In order to increase confidence in my results, and to facilitate the fastest possible comprehension of the data, I moved iteratively between an attempt to interpret the data and the two data-generating strategies chosen, semi-structured interviews and consultation of the literature. This retroductive research strategy allowed me to repeatedly formulate and then test descriptions of the state of the epistemic communities identified and their beliefs. These repeated tests allowed me to refine my descriptions so as to increase the practical adequacy with which they described the data. I had intended to cease to generate more data when these new data were consistently not providing any new challenges to my descriptions. This is the point of ‘data adequacy’ also referred to as ‘saturation’ (Morse, 1995). In reality this point is reached by degrees as small nuances can always be added to descriptions by new data. In this case, after conducting and interpreting 12 interviews my broad descriptions were no longer being challenged, though much nuance was still being contributed when interpreting the 11th and 12th transcripts. With the relatively small *N* of 12, it also felt possible to me that a significant challenge to my descriptions might be added by a new interviewee, but I had run out of time to arrange, conduct and interpret further interviews without further funding, and I was content that the probability of a major upset to my descriptive categories seemed small. The data generated are discussed in Chapter Eight.

4.3 ANSWERING THE PRIMARY RESEARCH QUESTION

> *Can we give a useful, systematic account of the relative merits of evidence generated using different (quasi-)experimental development impact evaluation methods that goes beyond internal validity to also consider the extent to which methods facilitate the transfer of results to other contexts? If so, how?*

The protocols outlined in Sections 4.1 and 4.2 of this chapter produce two sets of findings. The first protocol generates a quantitative scoring of different methods with respect to the extent that they tend to facilitate the transferability of results, and a set of qualitative insights into the ways in which transferability is facilitated by these different methods. The second protocol generates a description of the major anglophone epistemic communities that claim authority over the quality of research findings in the evaluation of development interventions; their approaches to evidence quality assessment; and the features of these accounts that are valued by members of the community as well as any unresolved puzzles or nascent crises that put pressure on them. Both protocols are a sort of mapping exercise, though a critical assessment of the terrain is also implied, in particular by the ranking generated by protocol one. These exercises provide the necessary data to generate an answer to the primary research question. In order to realise this potential, the final stage of research proper is an attempt to construct an account that satisfies the demands of the primary research question. That is, a systematic account of the relative merits of evidence generated using different methods that considers the facilitation of transferability as well as internal validity, and that is useful to development experts.

As Subsection 4.1.3 has identified, the use of a novel method necessitates splitting the operationalised interpretation of research subquestion one into two parts.

1.

a) Can realist programme theory mapping be adapted to create a tool to assess the transferability of (quasi-)experimental development impact evaluation results?

b) If so, what can it tell us about the relative merits of evidence generated using different methods, both

i. as they are currently used?

and

ii. as they might be used?

Therefore, the first step to interpreting the outputs of the protocol developed in response to this subquestion is to assess the success of the novel method employed. It must be established that the method employed is systematically informative, generating results that can be triangulated. It must be verified that they can be made sense of with reference to wider theory and can be corroborated with reference to the wider empirical literature. Chapter Seven relates this process, providing arguments that the method was successful. The second step in interpreting the outputs of the research protocol generated in response to the first research subquestion is to investigate what evidence has been generated relating to systematic differences between methods regarding the extent to which evaluations employing those methods facilitate the transfer of results to other contexts. Chapter Nine describes this process, relating the insights generated in response to 1.b)i. and ii. Methods are compared as they are currently used and some suggestions made for improvements to the practice of researchers employing various (quasi-)experimental impact evaluation methods.

It may seem strange that answers to research subquestion 1.a) are presented in Chapter Seven and answers to 1.b) are presented in Chapter Nine, two chapters later. This is because it is important to present results from the investigation of research subquestion 2. before the answer to 1.b). and the attempt to answer the primary research question, which are both presented in Chapter Nine. This presentation reflects the fact that the research subquestions were not worked on and answered in turn, but rather through an iterative, overlapping process. As discussed, the data generated by protocol two are used to shape the answer developed to the primary research question in such a way that its chances of being useful to development experts are maximised. This implies that the process of constructing the answer to the primary research question involves an iterative movement between the answer under development and the data generated by both protocols, as well as an engagement directly with the participants identified by protocol two. This deeply mixed-methods approach, featuring participatory involvement as well as methodological triangulation is informed by the schema in Mayoux (2006).

The attempt to construct an account that is 'useful' implies a dissemination process in which outputs beyond the thesis are produced. These outputs will certainly include journal articles, which are currently under development. In addition, it may be desirable to produce policy briefs, or other non-traditional outputs, such as an online mini-site to present the results in the clearest and most complete way possible. Such a dissemination will have to extend beyond the time allocated for writing up and require further funding, for example through a short-term contract with my former employer, the Center for Global Development, or as part of a post-doctoral fellowship. Both of these options are being pursued, though I am currently focussed on an application and extension of some of the ideas in this thesis through a research project funded by the Centre for Excellence in Development Impact and Learning, as discussed in Chapter Nine.

4.4 ETHICAL CONSIDERATIONS

The research methodology described in this chapter does not imply any affirmative answers to the questions included on the SOAS Upgrade Form Research Ethics Checklist. Nevertheless, compliance with the SOAS Research Ethics Policy and will require careful attention to some ethical issues. SOAS researcher commitments require me to gain the explicit, informed consent of participants to my research before conducting data-generating interviews with any of them. To achieve I produced a written consent form using the model template supplied to PhD candidates by SOAS. It is also my intention, though this is not required by the SOAS Research

Ethics Policy, to publish the data generated in response to research subquestion one as fully as possible in order to facilitate other researchers using it for replication or reanalysis. No commitment can be made to publish in full the data generated in response to research subquestion two, as this would involve publishing transcripts of conversations with participants in contravention of the confidentiality statement included on the SOAS consent form template with which I have complied.

Part two: lessons learned

5 Case One: conditional cash transfers for school enrolment

This chapter sets out in detail how the principles and methods described in Chapter Four were deployed in practice for the first case identified. Like the second case, this case is a pairing of intervention and outcome. The pairing identified is that of conditional cash transfers (CCTs) for school enrolment.

The method motivated and described in the first section of Chapter Four requires many decisions to be made by the researcher as they employ the method for a specific case. This chapter motivates and describes those decisions for the first case, as well as defending them against some anticipated criticisms. In addition, the process of employing the method described in Chapter Four generates considerable intermediate results in the form of the list of evaluations identified, the map of programme theory or theories that underpin those evaluations, and the determinants of transferability for the set of evaluations. These are of interest in their own right and must be understood in detail in order to understand the presentation of final results that follows in Chapters Seven and Nine. They are therefore presented here for readers. These intermediate results also embody many methodological judgements on the part of the researcher that must be presented transparently in order to be visible and interrogable for readers.

The chapter proceeds by presenting the methodological decisions required and intermediate results generated by each of four stages of the method described in Chapter Four. First, the sampling of evaluations is discussed. Section Two discusses the identification of the intervention theory or theories that underpin the evaluations in the set. The theory identified is expressed as a collection of context, mechanism, intervention, outcome configurations or CIMOs. Section Three discusses the movement from an understanding of the mechanisms claimed to be involved in intervention causation to the creation of a list of markers that evaluations need to report in order to facilitate an argument for the transferability of their results to any target context. Section Four discusses how that list of markers was used to generate the dataset of final results that is described and discussed in Chapters Seven and Nine.

5.1 IDENTIFYING THE SAMPLE OF EVALUATIONS

The most-studied intervention-outcome pair identified by analysis of the AidGrade database was conditional cash transfers (CCTs) and school enrolment, with 29 evaluations identified. In accordance with the research protocol outlined in Chapter Four, this observation determined the selection of the set of evaluations of the effect of CCTs on school enrolment as the first case to be examined.

The first step in examining this case was to identify more recent evaluations that could be added to the sample in order to ensure that modern, best-practice evaluations were included. As described in Chapter Four, these papers were selected from the 3ie (quasi-)experimental impact evaluations repository. Fortunately for this research project, all evaluations in this repository are tagged by intervention 'sector'. These tags include one for conditional cash transfer (CCT). Unfortunately for this research project, evaluations in the repository are not also tagged by outcome variable. Therefore, it was necessary to manually screen evaluations tagged as CCTs for the reporting of school enrolment as an outcome variable. Because the purpose of screening this database was to identify evaluations conducted since AidGrade's screening process was completed, only evaluations published during or after 2013 needed to be screened.

An initial search for evaluations tagged with 'conditional cash transfer' yielded a list of 256 short entries describing an evaluation. It was not possible to sort these entries by publication date because the web search does not permit this and 3ie were unwilling to share an electronic version of the database of evaluation meta data. Therefore, the list of 256 short entries was screened manually for date of publication, identifying 62 entries listing an evaluation with a publication date of 2013 or later. Abstracts of studies published up to 2013 were screened to identify any evaluations not included in the AidGrade database but which might be relevant. For the 62 evaluations published in 2013 or later, full texts were examined to ascertain whether school enrolment was a reported outcome variable. Screening out evaluations for which this was not the case identified 20 evaluations of relevance. Reconciling this list with the AidGrade database and screening out duplicates of the same evaluation either across the two databases or through different publications (e.g. an earlier working paper and a later journal article) led to the identification of 12 additional evaluations. In all cases, the latest published incarnation of the same evaluation was screened in and the older version screened out. The total sample of evaluations identified across both databases numbered 41. However, later full text analysis resulted in the disqualification of four evaluations in the set for being analyses not of CCTs but of UCTs, taking the total down to 37. Descriptive statistics for this set of evaluations are available in Table 5.1. A full list of the evaluations included in the set for case one is available in Appendix A.

Table 5.1: Descriptive statistics of the set of evaluations of CCTs for enrolment

Total	**AidGrade**	**3ie**	**RCT**	**Diff-in-diff**[20]	**PSM**[21]	**RDD**[22]	**IV**[23]	**2000-2009**	**After 2009**
37	26	11	22	9	4	3	2	12	25

N.B. The sum of the methods counts is 40, reflecting the fact that three evaluations in the set employed two methods.

5.2 IDENTIFYING THE INTERVENTION THEORY OR THEORIES

The next step in examining this case was to identify the model or models of intervention causation that predominate in the literature and to synthesise these where possible to create a realist interpretation of the programme theory or theories behind the set of evaluations. As Chapter Four, Subsection 4.1.3 describes, this process is similar to the creation of an initial programme theory during the early stages of realist synthesis. However, the objectives of the two exercises are subtly but importantly distinct. In realist synthesis, an initial programme theory is developed and is then extended and refined though an iterative movement between

[20] Difference-in-differences approach
[21] Propensity score matching approach
[22] Regression discontinuity design
[23] Instrumental variable approach

theory and evidence. This is a process of ‘identifying, testing out, and refining programme theories’ with the ultimate goal of explaining ‘what works for whom, in what circumstances, in what respects, and how?’ (Pawson et al., 2004, pp.v, 20). Programme theory, then, is refined in response to evidence of whatever sort, with the ultimate goal of better describing states of affairs in the world.

The purpose of the exercise described in this section is different. The aim is to describe the theory or theories of intervention causation that predominate in the texts of the set of evaluations that have been identified. The ultimate goal is a complete description of those theories as well as an assessment of which evaluations are grounded in which theory or theories. This allows the complete description to be refined to a more restrictive description of just those theories that have widespread support across the set of evaluations, and for the evaluations in the set to be divided into subsets that share the same theoretical underpinnings.

The ontological and epistemic approach of this research project are the same as those of practitioners of realist synthesis, so tools can be adapted from realist synthesis to aid with this part of this research project. However, the approach to literature reviewing cannot be copied wholesale, because, as discussed, the purpose of this part of this research project is different. The key tool to be adopted from the realist synthesis toolkit is that of ‘purposive sampling’ to a point of ‘theoretical saturation’ (Pawson and Tilley, 1997; Pawson et al., 2004).

5.2.1 Purposive sampling of the relevant literature

The end goal of this literature review was to uncover the programme theory or theories that underpin the set of evaluations identified. As described in Chapter Four, it was necessary to discover the theory that underpinned every evaluation in the set. This is because evaluations were to be compared with each other on the extent to which they reported the markers implied by the theory that underpinned each evaluation. Theories shared between evaluations could be aggregated and synthesised, but distinct theories only supported by one or two evaluations were a possible feature of evaluations in the set, and so this possibility had to be explored. As described in Chapter Four, a three-stage literature review strategy was adopted for both cases.

For case one, this proceeded as follows: First, all 37 evaluations were read closely in turn, and the information relating to programme theory was extracted and added to a document functioning as a theory map in progress. In the second stage, seven reviews identified through web searches were interrogated for theoretical information and their theoretical references followed until no new information was being added to the theory map in progress.

In the third stage, an evaluation from the set was selected at random. As described in Chapter Four, references relating to programme theory from this evaluation, the ‘root evaluation’, were collected to create a set of further theoretical references. The referenced documents were then read closely and all information relating to programme theory was extracted from them and used to update the theory map in progress. For any major points of theory that were referenced to additional documents, these were also added to the set of theoretical references to be followed. This extension could continue more than once, until the theoretical allusion of the root paper had been sufficiently elaborated to be considered well-grounded in the literature. Once this iterative branching search strategy had been exhausted, a different root evaluation was selected at random, and any new theoretical references followed and investigated. This retroductive movement between programme theory map and source documents was continued until no significant additions had been made to the programme theory map for three consecutive root evaluations. As Chapter Four, Subsection 4.1.6 describes, this is the point at which the theory map has achieved practical adequacy in its description of the theory underpinning evaluations in the set.

5.2.2 Generating the theory map

The literature review strategy described above resulted in the investigation of all 37 of the evaluation documents in the set, seven reviews, and then 25 theoretical references from 7 root evaluations before the theory map presented in Table 5.2 emerged and remained unchanged for long enough to be considered practically adequate. The theory map contains only theory of relevance to the causation of enrolment outcomes by CCTs of different kinds in specific contexts. The purpose of creating this theory map is to facilitate an investigation of the features of context and the features of the intervention that must be reported by an evaluation in order to facilitate a reasoned argument for the transferability of evaluation findings from the evaluation context to some target context. More detail on the inclusion/exclusion decisions that were made when generating the theory map is included in Subsection 5.2.3.1.

In accordance with the realist epistemology adopted for this research project, the theory map is presented as a collection of contextual features which combine with intervention features to activate mechanisms that lead to outcomes. Individual context-intervention-mechanism-outcome configurations (CIMOs) can be read across each row of the table. Sometimes a cell is merged across rows, representing e.g. a particular feature that is a constituent of more than one configuration.

It is more usual in the realist synthesis literature for programme theory to be represented as a collection of context-mechanism-outcome configurations (CMOs). When an intervention of a given type, e.g. a CCT, is implemented in a given setting, intervention features combine with features of the setting to create the total environment that is described as ‘context’ in the traditional CMO framework. However, this way of talking about features of setting and intervention features obscures relationships between the two. By separating the two, their relationships can be made clearer. For example, for CCTs, intra-household bargaining problems constitute the setting in which two distinct intervention features (making transfers to mothers and making transfers conditional) can activate two different mechanisms (the empowerment mechanism and the substitution mechanism), as Table 5.2 shows. The CIMO framework may first have been suggested by Denyer et al. (2008) and has been found useful by many authors of realist evaluations and syntheses.[24]

No ontological difference between contextual features and intervention features is implied by the use of the CIMO framework, nor any challenge intended to the core principle of realist social science research that the basic unit of causal theorising is a ‘context-mechanism-outcome

[24] See, for example, Mazzocato et al. (2010), Astbury and Leeuw (2010), Frykman et al. (2017) and Maidment et al. (2017)

combination' (Pawson and Tilley, 1997, p.220). Aspects of a programme do indeed combine with the causal powers and liabilities of objects, agents, institutions and all of the other aspects of a given place and time to create a context in which mechanisms are activated, interact with each other, and produce outcomes. Therefore, in reality, no intervention is ever implemented 'the same' in two different settings because of the way interventions must combine with setting in order to be realised (Pawson and Tilley, 1997, p.133). However, there are commonalities between different instances of the same sort of intervention. It is illuminating for the purposes of this research to separate these features of an intervention from the features of setting with which they combine in order to speak about commonalities and differences across settings. Therefore, Table 5.2 separates contextual features from intervention features, preferring to employ the CIMO framework over the CMO framework.

Table 5.2 is presented below, with each element from the table explained in the following section.

Table 5.2: Final programme theory map represented as CIMOs

<table>
<tr><th>Contextual feature</th><th>Intervention feature</th><th>Mechanism</th><th>Outcome</th></tr>
<tr><td>Financial barriers to education creating levels of enrolment below the perceived privately optimal level
• Available resources < direct costs + indirect costs of education[25] And imperfect credit markets[26]</td><td>Transfer of money to households</td><td>Income mechanism – changes available resources to reduce the liquidity constrains limiting households’ ability to invest in educating their children[27]</td><td>All children are more likely to be enrolled, but mediated by parents’ tendency to want to enrol them</td></tr>
<tr><td rowspan="2">Intra-household bargaining problems creating levels of enrolment below the mother’s perception of the privately optimal level for some or all children in the household
• Excessive future discounting on the part of one or both parents towards one or both sexes and imperfect credit markets[28]
• Other intra-household factors driving differing perceptions of the privately optimal level of enrolment.</td><td>Transfer to mothers</td><td>Empowerment mechanism – increases the decision-making power of mothers in order to leverage the fact that their preferences are more closely aligned than fathers’ with children’s interests, perhaps especially the interests of ‘marginal’ children.[29]</td><td>The children negatively affected by intra-household bargaining problems are more likely to be enrolled</td></tr>
<tr><td rowspan="2">Conditionality attached to transfers (with some degree of enforcement)</td><td rowspan="2">Substitution (or price) mechanism – increases the expected cost of not educating children by a fixed amount including for ‘marginal’ children[30]</td><td rowspan="2">All children are more likely to be enrolled, and the moderating effect of parents’ tendency to want to enrol is reduced, increasing enrolment rates for marginal children by more than privileged children</td></tr>
<tr><td>Misguided beliefs creating levels of enrolment below the true privately optimal level (and/or below the socially optimal level)
• Absent or erroneously low information about returns to education, other erroneous beliefs, or damaging familial or community norms.
○ Perhaps acute for ‘marginal’ children[31]</td></tr>
</table>

N.B. The difference in vertical alignment is deliberate between the contextual feature ‘Misinformation or misguided beliefs…’ etc. and the cells to its right. This represents the fact that conditionality, the substitution mechanism, and the change in outcomes that these cause are all enabled both by misguided beliefs and by intra-household bargaining problems whereas transferring resources to mothers, the empowerment mechanism, and the changes in outcomes this causes is not linked to misguided beliefs, but only to intra-household bargaining problems.

[25] Universally described. E.g. Snilstveit et al. (Snilstveit et al., 2015, p.138) and even highly theory-averse treatments such as Conn (2017, p.73). Edmonds (2007) discusses the opportunity cost of schooling in lost child wages.

[26] Fiszbein and Schady (2009, p.56), Angelucci et al. (2010) discuss the extent to which an active extended family network can stand in for credit markets in some contexts.

[27] Amarante, Ferrando and Vigorito (2013), DeBrauw et al. (2011, pp.312–313) also discuss changes in time allocation that may lead to more time being available for schooling. If time is considered a resource, then this insight can be subsumed under the ‘income’ and ‘price’ effects.

[28] Benhassine et al. (2013), Fiszbein and Schady (2009, p.58)

[29] Baird et al (2013, p.2)

[30] Baird et al. (2011), Benedetti et al. (2016)

[31] Akresh *et al.* (2012a; b; 2013)

5.2.3 Presenting the elements of the theory map

This section provides more detail on the CIMOs described briefly in Table 5.2. First, however, elements of theory that were considered and excluded are discussed.

5.2.3.1 Elements considered and excluded

The elements of theory presented in this subsection were all present in the literature reviewed but were not included in the final programme theory map. These exclusions are motivated by a variety of reasons which are discussed in more detail below. What all these reasons have in common is that they reflect a way in which the theoretical element considered would not have contributed to a model of programme theory that served the role required of it for the purposes of this research project. As Chapter Four described, the purpose of the theory map is to permit the assessment and comparison of the evaluations in the set as regards the manner and extent to which they assess and report on those markers of context and intervention implementation of relevance to an argument for the transferability of evaluation findings. All of the theory elements below would have undermined the programme theory map's ability to fulfil that role, for a variety of reasons.

5.2.3.1.1 Downstream and upstream effects

The theory map presented in Table 5.2 is limited to theory of the causation of the outcome by the intervention. In this case, that means that the only relevant programme theory was theory relating to the causation of increased enrolment by CCTs. In realist terms, this means theory about the mechanisms that connect intervention to outcome via contextual features and features of intervention implementation. This means that downstream effects of a change in the outcome, e.g. increased human capital and earning potential, were not relevant objects of programme theory. Also not relevant for the purposes of this research project were upstream theories about the way in which contextual features arose, except where those mechanisms were the target of intervention features. So, the fact that reaching a 'transition grade' reduces levels of enrolment is not relevant to the propensity for a CCT to increase enrolment by some proportion because no feature of the intervention acts to change this mechanism active in the context (Schady and

Araujo, 2008, pp.58–59). However, the fact that low levels of enrolment can be caused by parents' perception that it is not worth sending lower ability children to school is a relevant programme theory insight. This is because CCTs aim, through the mechanism of a substitution or price effect, to increase the opportunity cost of not educating children, making parents more likely to choose to enrol lower ability children (Akresh, De Walque and Kazianga, 2013).

5.2.3.1.2 Effects of the intervention on non-recipients

Ferreira et al. (2009, p.2), propose the existence of a 'displacement effect' in which 'cash transfer programs conditional on the school enrolment of one specific child might lead parents to reallocate child work away from the recipient and to other children in the household.' They go on to say that '[m]ore generally, the transfer may provide an incentive for parents to specialize in the education of the recipient, leading to a displacement of—less schooling for—his or her siblings.' Barrera-Osorio *et al.* (2008) find evidence for this effect in Colombia. This possible mechanism acts only on children external to the recipients of the CCT. Detecting such externalities, or 'spillover effects', requires a different evaluation design and a different set of theoretical understandings such as that proposed by Barrera-Osorio (*ibid*). As the goal of this theory-mapping exercise is ultimately to provide a means of comparing evaluations, and as very few evaluations in the sample are designed to detect spillovers, the scope of this assessment was limited to the effect on outcomes for recipients and theorising about spillovers was excluded from the theory-mapping exercise.

5.2.3.1.3 Contextual factors that justify the form of the intervention but do not determine effectiveness

The contextual factors that *justify* the form of the intervention but do not determine effectiveness were also not included in the programme theory map. In the case of CCTs, there is a large literature regarding the conditions under which the conditionality of the transfer is warranted.[32] So, for example, the existence of an anti-poor political economy might justify designing an intervention as a CCT rather than an unconditional cash transfer (UCT), even if the

[32] See, for example, Das *et al.* (2005), Fiszbein and Schady (2009) and Baird, McIntosh and Özler (2011).

unconditional cash transfer would promote the desired outcomes at a lower cost. This is because the CCT, by tying in to narratives about the need to distinguish the 'deserving and undeserving poor', would be more likely to be implemented and to survive a change of government. However, this theory says nothing about the causal effect of the conditionality on the outcome of interest for this research project (or any outcome). So, this theory was not a candidate for inclusion on the programme theory map constructed in the course of this literature review. By contrast, the existence of widespread incorrect beliefs about the returns to education in the study population justifies implementing a CCT over a UCT by making the CCT likely to be more effective than the UCT. Similarly, the existence of an average level of investment below the true privately optimal level but at the perceived privately optimal level both justifies conditionality and determines a lower level of effectiveness for the intervention by ensuring that the income mechanism contributes little to changes in enrolment. These elements of programme theory therefore were a candidate for inclusion on the programme theory map and are represented in Table 5.2.

5.2.3.1.4 Non-essential intervention features

Many of the evaluations and theoretical references examined in the course of the theory mapping process discussed the effect of mechanisms activated by non-essential features of a conditional cash transfer intervention. For example, Kabeer *et al.* (2012, p.8) discuss the 'training of various kinds' that sometimes accompanies a CCT. However, such training is not an essential feature of a CCT. What is meant by this is that training can be and usually is not included in a CCT intervention. Training is a bolt-on feature, sometimes added, not a feature without which the CCT ceases to be a CCT. This is in contrast to some sort of conditionality, without which a CCT ceases to be a CCT, and becomes a UCT. The purpose of the theory map is to facilitate the comparison of the evaluations in the set. In order to maximise the sample size for such comparisons, it is desirable not to split the evaluations in the set into subsets that employ different theories and therefore should report on different contextual and intervention features, unless this is necessary. One way in which this can be avoided is to restrict the analysis to the causation of the outcome by the essential features of the intervention. That is why the conditionality mechanism is included in Table 5.2, but no mechanisms associated with training are. An evaluation of an intervention that includes a training component alongside its CCT can still be assessed and compared with other evaluations on the extent to which it reports on the contextual and intervention features necessary to assess the functioning of the mechanisms activated by essential intervention components. This analysis leaves to one side the extent to which this evaluation also reports on the contextual and intervention features necessary to assess the functioning of the mechanisms associated with training, but such an assessment is not necessary for the purposes of this research project.

5.2.3.1.5 Mechanisms without widespread support

The literature examined is remarkably homogeneous with regards to programme theory, with some differences of emphasis but very little disagreement about the existence of essential mechanisms. The small amount of disagreement that does exist is reflected in two mechanisms that were considered for inclusion but not included. The first of these appears in Davis *et al.* (2016). They posit the existence of a 'hope' mechanism whereby 'cash transfers make people happier and give beneficiaries hope, a precondition for families to want to invest in the future.' They provide evidence from six evaluations of CCTs in Ghana, Kenya, Malawi, Zambia and Zimbabwe that were designed to detect changes in 'happiness', 'quality of life', 'belief that life would be better both two and three years in the future', 'feeling that [households] are now better off' and 'satisfaction with life', respectively (*ibid,* p.337).

The second proposed mechanism not included is the 'performance' mechanism suggested by Dubois et al. (2012). They propose that the conditionality of a CCT creates an incentive to attend school which may raise performance. This performance may reduce the chance of dropout in situations where continued enrolment is conditional on a passing grade or where the student would otherwise be forced to repeat the year. Dubois et al. claim that descriptive statistics of households receiving Progressa/Opportunidades bear out a positive relationship between repeating the year and dropping out.

For both of these mechanisms, the evidence presented and the reasoning behind this proposed mechanism are somewhat compelling, but the mechanisms do not have widespread support in the literature; they were not discussed by any of the other sources consulted. Davis *et al.* (2016) is a review of available evidence from SSA. None of the evaluations in the set were designed or reported in a manner motivated by the existence of the hope mechanism suggested by Davis *et al.* (*ibid*). Therefore, this mechanism must be excluded when assessing the extent to which evaluations in the set explore and report on the contextual and intervention features of relevance to an argument for the transferability of their findings. Dubois *et al.* (2012) is an evaluation in the set. However, the mechanism that they propose is only discussed in two other evaluations in the set, and not in the major theoretical works cited by many evaluations. This divergent theoretical understanding could be thought to motivate splitting the set into two subsets, a larger set, and a smaller set containing the three evaluations including Dubois *et al.* that talk about this mechanism. However, this was not warranted. Dubois *et al.* and the other two evaluations that could have been included in this small subset all share with the other evaluations the theoretical understanding represented in Table 5.2. While they posit the existence of a further mechanism, this is additional and not contradictory to the existence of the mechanism in Table 5.2. Because this mechanism is additional, it is legitimate to compare these three evaluations with the rest of the evaluations in the superset regarding the reporting of intervention and contextual features of relevance to the action of the three mechanisms in Table 5.2. The three evaluations could additionally be compared with each other on the reporting of intervention and contextual features of relevance to the action of the additional mechanism they posit, but this comparison is of such a small sample of similar evaluations that it was not considered informative and is not reported here.

5.2.3.1.6 A newly identified mechanism

It is compellingly argued by Gaarder (2012) and Baird *et al.* (2013), that cash transfer programmes exist on a continuum of conditionality from unconditional cash transfers with no explicit associations with e.g. health or education to conditional cash transfers with conditions that are known to recipients, well monitored and enforced. In between these two extremes are labelled transfers that are associated to some extent with use for a particular purpose, either through their name, the location of disbursement or even an explicit set of directions for use or a stated conditionality, but without any enforced conditionality. Benhassine et al. (2013) have even tested the strength of the labelling mechanism in one context, finding almost no difference on that occasion between a labelled transfer (but with no stated conditionality) and an enforced conditional transfer.

If a labelling mechanism might explain most of the effectiveness of conditionality in some contexts, then it would seem important to include this mechanism in the theory map generated for this research project. However, Gaarder's (2012) commentary in the *Journal of Development Effectiveness* is the first time that this mechanism is hinted at in the literature. Only 7 CrossRef citations and 15 Google Scholar citations are recorded for this article. It is not until Baird *et al.* (2013) that a high-profile paper (197 Google Scholar citations to date), suggests the operation of this mechanism. Of the 41 evaluations included in the sample for this case study, 25 are published in 2012 or earlier, with another 7 published in 2013. The labelling mechanism and its associated CIMO therefore could not be included in Table 5.2 or the analysis that follows. This is because it is clearly not a mechanism that can be considered to be present in the theoretical understandings evaluations in the set other than Baird *et al.* (2013) and Benhassine *et al.* (2013). Only nine evaluations in the set were published after the publicising of this mechanism, and none of the other seven mention it. As this is an additional rather than contradictory mechanism, just as with the performance mechanism discussed in the previous subsection, Baird *et al.* (2013) and Benhassine *et al.* (2013) can be included in the main set of evaluations and need not be compared separately.

5.2.3.1.7 Reasonably implicit contextual features

Maluccio *et al.* (2010) and others in the literature correctly point out the important role of supply side constraints in conditioning the effectiveness of CCTs aiming for improvements in school attainment. Of course, this theory map is only concerned with theory concerning how and the circumstances under which CCTs might be effective at increasing *enrolment*. However, the supply side might still be reasonably argued to be important. It will not be possible for CCTs to increase enrolment if there are no schools available or if schools refuse to enrol additional pupils. However, this contextual feature can reasonably be taken to be implicit for evaluations concerned with enrolment. It would not be reasonable to penalise an evaluation for failing to report the contextual information necessary to facilitate a reasoned argument for the transferability of findings about the effectiveness of a CCT on enrolment rate because it failed to report the fact that schools willing to enrol students were present. The same is true for the presence of imperfect credit markets in the evaluation context. While this is necessary for the action of the income mechanism, imperfect credit markets are a feature of all of the settings where conditional cash transfers might reasonably be evaluated, and therefore their presence does not need to be reported as such. Reporting the geographical area in which the evaluation took place, as all evaluations in the set do, can be taken to be sufficient for the reporting of this contextual feature.

In practice, including such reasonably implicit contextual features in the list of contextual features to be reported against which the evaluations are scored would, in expectation, have the effect of increasing average score and adding noise to the comparison. Average scores would increase because almost all evaluations would report something equivalent to the existence of these contextual features. Those that didn’t would be likely to have done so out of a reasonable belief that the information was implicitly included. Judging evaluations on the existence or absence of such a belief on the part of the evaluator(s) would add a random element to the following comparison rather than rendering it more informative. Contextual factors that might reasonably be taken to be implicit were therefore excluded from the theory mapping process.

5.2.3.2 CIMOs organised by mechanism

This subsection provides a more detailed account of the three CIMOs summarised in Table 5.2, explaining the provenance of those theories through references to sources. It also facilitates an argument for the identification of all and only those contextual features and intervention features that are identified in the following section as necessary for an argument for the transferability of evaluation findings. As the previous subsection mentioned, the literature sampled was remarkably homogeneous in its theorising about mechanisms. The three CIMOs elaborated upon below have such widespread support across the literature that almost every evaluation either discusses them at some length or mentions them briefly and references one of the key theoretical works such as Fiszbein and Schady (2009).

5.2.3.2.1 The income mechanism

The most basic manner in which CCTs combine with contextual features to produce increased enrolment is through an income mechanism. The activating of this mechanism by the intervention, in the context, leading to a change in outcome is referred to in the economics-dominated literature as 'the income effect'. This CIMO requires the presence of a context in which households face financial barriers to investing as much as they would otherwise choose into the education of their children. In order for this contextual feature to hold, households must be liquidity constrained by the absence or poor functioning of credit markets (Fiszbein and Schady, 2009). They must also face financial barriers specifically to education in the form of direct costs such as school fees, uniforms, transportation, etc. and/or indirect costs in the form of returns to activities that are constrained by enrolment in school, such as child labour (Edmonds, 2007; Snilstveit et al., 2015).[33] Households who would otherwise not face financial barriers to education may also be savings constrained due to the costs of enrolment in school being concentrated at one time in the year as well as imperfect savings institutions or commitment issues that prevent adequate saving (Barrera-Osorio et al., 2008). However, savings constraints are rarely discussed in the literature and so this contextual feature has not been included in the programme theory map.

In the presence of financial barriers to enrolment the context will be characterised by a level of enrolment that is below what the literature refers to as the household's 'privately optimal level'. By this is meant that level of enrolment that the household would choose, were that choice not

[33] It is not correct to say that child labour is ruled out by school attendance, let alone enrolment. This is especially true in a context such as Brazil where children attend school in four-hour shifts and often work around this. However, schooling at the very least constrains this activity.

constrained. However, as many in the literature point out, there is a further distinction to be made between the ‘true’ privately optimal level of investment in education (and therefore school enrolment), and the level that is perceived to be optimal by households.[34] The ‘true’ privately optimal level of investment in education for a given agent, i.e. an individual or a household, is that level of investment that maximises rationally expected returns. There will be a difference between this level and the level that is perceived to be optimal by the agent if the agent suffers from failures of rationality or holds erroneously low beliefs about the returns to education. It will also diverge if the agent is optimising on dimensions other than future returns e.g. in the presence of familial or community norms about what is an appropriate level of education for whom. It is this perceived privately optimal level of enrolment that is important for household decision-making, rather than the ‘true’ level. In order for the income mechanism to activate, there must be households in the population targeted by the intervention who are prevented from achieving their perceived privately optimal level of enrolment by financial barriers.

The only relevant intervention feature to the operating of this mechanism is the transfer of money to households. The transfer changes the resources available to households in order to ease the liquidity constraints limiting their ability to invest in educating children. The resultant outcome is that all children are more likely to be enrolled, but this is mediated by parents’ tendency to want to enrol them. The transfer will be more effective in activating this mechanism if it is larger and if it is more regular and predictable (Davis et al., 2016, p.344).

It should be noted that this mechanism also operates for unconditional cash transfers. This mechanism will not result in higher rates of enrolment for children for whom the expected returns to education do not outweigh the costs of education. It is the existence of a gap between the perceived private optimal level of enrolment and the real private optimal level and/or the socially optimal level that motivates the use of a CCT over a UCT. This is discussed in more detail in Subsection 5.2.3.2.3, below.

[34] Fiszbein and Schady (2009) is the most complete treatment.

5.2.3.2.2 The empowerment mechanism

Another mechanism shared with UCTs is the empowerment mechanism that can be activated by targeting mothers as the beneficiaries of the transfer rather than the head of household, likely to be a man. This mechanism is activated when intra-household bargaining problems create a level of enrolment below their perceived privately optimal level for children, either all children or those who are not favoured by either one or both parents.

There is some disagreement in the literature about whether cultural norms and/or excessive discounting of future returns drive intra-household bargaining problems. Fiszbein and Schady (2009, p.9) characterise excessive future discounting by parents on behalf of one or both sexes as *the* driver of intra-household bargaining problems, which they call ‘incomplete altruism’ and a classic principal-agent problem. In the presence of perfect credit markets, these discount rates would not change investment decisions, but in their absence they will lead to lower enrolment (*ibid*). Baird *et al.* (2013, pp.10–11) separate excessive discounting from intra-household bargaining problems, characterising the latter as driven by ‘for example, parents not valuing girls’ education or community norms keep families from sending girls to school.’ In fact, intra-household bargaining problems are of course only caused by differences in the preferences of family members. Community norms will only drive bargaining problems within the household if members of the household accord them different levels of priority. For example, if fathers do not value girls’ education, this only creates an intra-household bargaining problem if other members of the household value girls’ education more highly. The same is true for excessive discount rates on the future returns to education. It is important to differentiate between intra-household bargaining problems and other drivers of lower levels of enrolment. This is because the empowerment mechanism will be activated only in the presence of the former and not the latter.

Transfers to mothers may activate an empowerment mechanism in situations where mothers do not share the excessive discount rate of the head of household. Transfers to mothers have been a feature of CCTs in the Global South since Opportunidades began, based on prior evidence from other populations that suggested mothers were more likely to spend transfers on investments in health and education than fathers (Das, Do and Özler, 2005). Activation of this mechanism should result in higher rates of enrolment for the children affected by intra-household bargaining problems.

Baird *et al.* (2010, 2013, p.2) argue that if at least one adult in a household attaches a lower value to at least one child's education than the child themselves, and possibly other members of the household, then transfers to the child might activate an empowerment mechanism by 'sticking' to them. It is very rare in the literature to transfer resources directly to children, however, so only transfers to mothers are described as a channel for the activation of the empowerment mechanism in the programme theory map.

5.2.3.2.3 The substitution mechanism

As mentioned above, CCTs as a means of boosting school enrolment are generally justified over UCTs by a gap between the perceived privately optimal level of enrolment and the 'true' privately optimal level. This may be due to absent or incorrect information about the returns to education, perhaps acute for low (perceived) ability children such as genuinely less able children, younger children, girls, or otherwise 'marginal' children (Akresh, De Walque and Kazianga, 2013). Fiszbein and Schady (2009, p.53) usefully point out that incorrect beliefs might also be about how human capital accumulates rather than about the returns to it, for example the belief that formal schooling requires high levels of natural talent not found in one's own household. However, whether erroneous beliefs are about processes of accumulation or about returns, the result is the same: a difference between the economists' expected return to households' children's education and the returns as predicted by those households. Levels of enrolment below the 'true' privately optimal level may also be driven by the excessive future discounting or familial/community norms discussed in the previous section, whether these also cause intra-household bargaining problems or not.

In order to motivate households to enrol their children in school despite the non-financial barriers to education discussed in the previous paragraph, CCTs employ the conditionality of their transfers to create a substitution or ‘price’ mechanism, which increases the expected opportunity cost of not educating a child. This cost is flat across all children, meaning that even unfavoured children, whose education may be valued at a low level by the decision maker(s) within a household, are more likely to be enrolled in school in order to qualify the household for the cash transfer conditioned upon their enrolment and usually their attendance above some threshold. The outcome of the operation of this mechanism is that all children are more likely to be enrolled, and the moderating effect of parents’ tendency to want to enrol is reduced, increasing enrolment rates for marginal children by much more than privileged children.

As discussed in Subsection 5.2.3.1.6, CCT interventions exist on a continuum of degrees of conditionality. Some programmes communicate conditions which are not monitored or enforced, others monitor conditions but these are only enforced to a degree, still others monitor and enforce conditions strictly. What is important for the substitution mechanism to activate is that households understand conditionality and expect conditions to be enforced. This does not mean that only perceptions of the probability of enforcement at baseline are relevant to an argument for transferability of findings; how well-monitored and enforced conditions are throughout the duration of an evaluation period will feed back to the intervention population and affect this expectation in later time periods. Therefore, the extent to which conditions are announced, monitored and enforced will be relevant to the extent to which this mechanism can be judged to have been activated.

The substitution mechanism will also operate where levels of enrolment are already at the true privately optimal level of the decision maker(s) but below 100%. Conditionality may still be desirable in these circumstances from the point of view of a policymaker because enrolment levels are judged to be below the ‘socially optimal level’. By this is meant that there are positive externalities and therefore benefits to society to enrolling children in school, and that these are not internalised in the decision-making of households. The substitution mechanism should still be activated and increase enrolment in these circumstances, as the price of not enrolling children will still apply (Das, Do and Özler, 2005).

5.3 FROM PROGRAMME THEORY TO DETERMINANTS OF TRANSFERABILITY

The previous section identified the principal mechanisms responsible for the change in enrolment due to CCTs. It is worth reiterating that the ultimate goal of this aspect of this research project is to use this case study to examine the systemic differences, if any, between (quasi-)experimental impact evaluation methods regarding the extent to which they facilitate the transfer of their findings to some other context. Chapter Three argued for the validity and the utility of the tools of realism to answer this question and gave it a theoretically rich interpretation:

> *What are the systematic differences, if any, between (quasi-)experimental impact evaluation methods regarding the extent to which they explore and report on the barriers and enablers of intervention mechanisms present in the study context and the extent to which different intervention mechanisms have been activated?*

The way in which this has been investigated is through the examination of a set of evaluations employing diverse (quasi-)experimental impact evaluation methods. Each evaluation needed to be assessed on the extent to which it reported on 'the barriers and enablers of intervention mechanisms present in the study context and the extent to which different intervention mechanisms have been activated.' These assessments were then compared within and between (quasi-)experimental impact evaluation methods in order to inform an answer to the secondary research question above, and ultimately this research project's primary research question.

The next step in the analysis of this case, then, was to move from the programme theory map given in Section 5.2 to an enumeration of the information that would have to be reported by evaluations of the effect of CCTs on enrolment rate in order to facilitate an assessment of 'the barriers and enablers of intervention mechanisms present in the study context and the extent to which different intervention mechanisms have been activated.' This was a two-stage process that began with an assessment of the barriers and enablers of intervention mechanisms and concluded with an assessment of the information required to determine the extent to which different intervention mechanisms had been activated.

5.3.1 Barriers and enablers of intervention mechanisms

An examination of Table 5.2 makes clear that the set of barriers and enablers of intervention mechanisms can be usefully broken down into a set of contextual features that constitute barriers and enablers and one of intervention features. Despite the fact that authors in the literature do not present straightforward lists of the contextual barriers and enablers of the mechanisms active in CCTs, these can be fairly uncontroversially inferred from the descriptions of programme theory described in the literature and synthesised above. Discussions of the intervention characteristics that constitute barriers or enablers to mechanism operation are more common in the literature as much of the theoretical literature is dedicated to discussions about the effect of intervention design choices on the effectiveness of the intervention. Therefore, the set of intervention features that constitute barriers and enablers to intervention mechanisms can be extracted from the CCT literature directly, despite not being described in the programme theory map above. Which such features are relevant to the activation of each mechanism was discussed in the previous section.

Table 5.3 lists contextual and intervention barriers and enablers by mechanism, using + and – to denote, respectively, whether increases in the listed parameter should be expected to facilitate or frustrate the action of the mechanism. Only those barriers and enablers that are implied by the programme theory described in Table 5.2 and Section 5.2.3 are included here. So, for example, despite the fact that Armand and Carneiro (2018) convincingly argue for the importance of the timing of payments through the school year as a determinant of programme effectiveness, this intervention enabler is not included because it relates to the importance of savings constraints in driving the income mechanism. Section 5.2.3.2.1 has explained that this element of programme theory does not have widespread support in the literature and so cannot be included for the purposes of comparing the extent to which evaluations explore and report on the contextual features and intervention features of relevance for an argument for the transferability of findings.

Table 5.3: Barriers to and enablers of the operation of intervention mechanisms

Mechanism	**Contextual barriers/enablers**		**Intervention barriers/enablers**	
Income mechanism	Difference between baseline enrolment level and perceived privately optimal level + • Perceived privately optimal level of enrolment + • AND Baseline level of enrolment -	Financial barriers to enrolment -[35] • Household available resources -[36] • AND Direct costs of education +[37] • AND Indirect costs of education + Non-financial barriers to enrolment - • Erroneously low estimates of expected returns to education - • Failures of rationality such as excessive future discounting - • Familial or community norms that restrict enrolment -		Size of transfer + Regularity of transfer +[38] Implementing institution(s) (and roles) +/-[39] Targeting criteria and method +/-[40]
Substitution mechanism	Difference between perceived privately optimal level and 100% enrolment level. + • Perceived privately optimal level of enrolment -[41]		Conditions announced +/- Monitoring + Enforcement through sanctions +	
Empowerment mechanism	Enrolment preference or rationality differences between household members +[42]		Transfer recipient +/-	

[35] While financial barriers to enrolment constitute a barrier to the action of the income effect mechanism, some degree of financial barrier is necessary for this mechanism to activate. The fact that Table 5.3 is not able to represent this threshold effect is an acknowledged limitation of this way of representing the set of barriers and enablers.

[36] While mean consumption would allow for basic comparisons across contexts, other features of the distribution are also desirable, as would information about the presence or absence of shocks such as a windfall harvest or drought (Gitter and Barham, 2009). This is true for many of the features discussed. However, reporting of means will be considered sufficient for the feature to have been reported as assessing the extent of the reporting of a potentially richer set of information describing of the distribution in a way that is comparable across studies is not feasible.

[37] In Akresh *et al.*(2013) the reported figure is ‘mean per child education expenses’. It is unclear from the paper whether this is per enrolled child or per child regardless of enrolment.

[38] Davis *et al.* (2016, p.344)

[39] Grosh *et al.* (2008, p.105) De Janvry et al. (2010) demonstrate the importance of the implementing institution and its incentives through showing how institutional setup matters for targeting, disbursement and enforcement of conditions.

[40] Fiszbeing and Schady (2009, p.173), Grosh et al. (2008, p.105)

[41] The ‘true’ returns to education are not relevant; they justify the intervention but do not determine its effectiveness. This is because none of the mechanisms identified attempt to communicate information about the ‘true’ returns. The particular drivers of the low level of enrolment are also not relevant.

[42] Could be measured directly or indirectly ex-ante or through outcomes ex-post.

5.3.2 Disaggregated outcomes as evidence of the action of mechanisms

Analysis of the CIMOs reported in Table 5.2 suggests that a limited assessment of the likelihood that different intervention mechanisms have been activated in a given study population can be derived from examining the barriers and enablers present. For example, if for a given study population households did not appear at baseline to be liquidity constrained from enrolling children in school, the income mechanism is unlikely to have been responsible for any change in outcomes.[43] This sort of evidence may be powerfully indicative that one or two mechanisms are more plausible candidates to have driven changes in outcomes. However, further examining the CIMOs reported in Table 5.2 suggests that disaggregated outcome variable data can provide an alternative set of premises for any such argument by enabling quantitative estimates of the relative importance of different mechanisms.

For example, if a CCT involving transfers to mothers has resulted in large enrolment rate increases, and these increases are 80% due to male-headed households 'catching up' to the enrolment levels seen in female-headed households at baseline, this will be powerful evidence that the empowerment mechanism has dominated other possible mechanisms as the source of enrolment rate increases. Therefore, outcome variable results disaggregated by gender of household head at baseline and endline are a powerful means of facilitating the assessment of the relative importance of the empowerment mechanism. Similarly, disaggregating outcome variable measurements by household wealth and/or consumption and by 'marginality' of child – for example based on gender, ability, or sibling age rank – will facilitate a compelling argument for the relative importance of the income and substitution mechanisms, respectively. Therefore, the reporting of outcome variables disaggregated on these three axes is an important component of the facilitation of the transferability of results.

The research protocol outlined in Chapter Four calls for the qualitative and quantitative comparison of evaluations in the case study set regarding the extent to which they 'explore and

[43] This case points out a limitation of the representation of barriers and enablers in Table 5.3 as discussed in footnote 34. While financial barriers to enrolment will frustrate the action of the income effect progressively as they increase, the existence of a minimum level of financial barriers to education is necessary for this mechanism to activate.

report on the barriers and enablers of intervention mechanisms present in the study context and the extent to which different intervention mechanisms have been activated.' This section has provided a list of the barriers and enablers of intervention mechanisms as well as three axes on which outcome variables should be disaggregated in order to facilitate an argument for the extent to which changes in outcome variable can be attributed to the action of each of the three mechanisms identified. Armed with this information, the evaluations in the set could be analysed qualitatively and quantitatively on the nature and extent of their reporting of these barriers, enablers and outcome measures.

5.4 GENERATING A QUANTITATIVE AND QUALITATIVE ASSESSMENT OF THE FACILITATION OF TRANSFERABILITY

The previous section has presented an argument for the existence of a set of barriers and enablers of intervention causation as well as three axes of disaggregation of outcomes which should be reported by an evaluation in the set in order to facilitate an argument for the transferability of results to some other context. These barriers, enablers and axes of disaggregation can be simplified to generate a list of causal markers. The term 'causal markers' is adapted from Cartwright (2019), who introduces it as a generic term for the visible signs of causally relevant underlying structures. To clarify what these are markers of, I also refer to them as MICCs, or markers of intervention causation in context. The reporting of these markers can then be assessed for each evaluation in the set in order to facilitate the comparison of evaluations which is necessary to answer the research question addressed by this strand of research. The full list of markers against which evaluations in the set were assessed is reproduced below for clarity:

Table 5.4: CCT MICCs

Group	**Subgroup**	**Marker**
Context	Level of enrolment	Household perceived privately optimal level of enrolment
		Enrolment preferences and/or rationality measures disaggregated by HH member
		Baseline enrolment
	Financial barriers	Household available resources
		Direct costs of education
		Indirect costs of education
	Non-financial barriers	Erroneously low estimates of expected returns
		Failures of rationality - excessive future discounting etc.
		Familial or community norms
Intervention	Conditionality	Announcement of conditions
		Level of monitoring
		Enforcement of sanctions
	Transfer	Transfer recipient
		Size of transfer
		Regularity of transfer
		Implementing institution(s)
		Targeting criteria and method
Outcomes		Enrolment by HHH gender
		Enrolment by HH wealth/consumption
		Enrolment by marginality of child

Each evaluation in the set was examined and scored 1 if it reported a given marker, 0 if it did not. In addition to the score variable, qualitative assessments of the reporting were also collected as a separate variable for each marker. In addition to this information, the ability of an evaluation team to generate data was also assessed and recorded as either low, medium, high or very high along with a qualitative explanation of why the score had been given. The use of four categories for ability to generate data was established in piloting as a satisfactory modelling simplification with adequate power to express differences between the evaluations in the set. 'Low' was recorded for evaluations that relied on secondary reinterpretation of already existing data, for example administrative data. 'High' was recorded for evaluations that made use of custom baseline and endline surveys. Most evaluations (20 and 12, respectively) fell into one of these categories. 'Medium' was recorded for three evaluations that fell somewhere in between. In one case, a custom survey was conducted only at the municipal level, and otherwise school records and administrative data were used. In another case, program eligibility score data was combined with a custom survey of a sample of households. In the final case, administrative records at baseline were combined with a custom endline survey. 'Very high' was recorded for two evaluations, which used custom surveys at baseline and endline that included household and individual-level units. This method of assessing the evaluations in the set resulted in a dataset with 37 observations and 42 variables, creating 1,554 data entries. The full dataset is available at https://mattjudendotcom.files.wordpress.com/2021/04/mjuden_thesis_data_and_code.zip and has been uploaded to the UK Data Service where it has been deposited in accordance with the conditions of my ESRC studentship under Project ID 204604.

The final results of the quantitative and qualitative assessment of evaluations in the set are reported in detail and discussed in Chapters Sevens and Nine.

6 Case Two: deworming for weight

Like the previous chapter, this chapter motivates and describes the methodological decisions that were made in the application of the method described in Chapter Four to a specific case of a pairing of intervention and outcome. The second such pairing identified and therefore the pairing discussed here is that of the administration of anti-helminth or deworming drugs in order to increase child weight gain.

As with the previous chapter, as well as motivating and discussing methodological decisions made, this chapter presents the intermediate results generated for the case at hand. These intermediate results take the form of a description of the set of evaluations identified, a map of the programme theory or theories that underpin those evaluations, and a list of the markers of intervention causation in context (MICCs) that must be reported by each evaluation in order to facilitate an argument for the transferability of results to some target context.

The chapter proceeds by presenting the methodological decisions required and intermediate results generated by each of four stages of the method described in Chapter Four. First, the sampling of evaluations is discussed. Section Two discusses the identification of the intervention theory or theories that underpin the evaluations in the set. The theory identified is expressed as a collection of context, mechanism, intervention, outcome configurations or CIMOs. Section Three discusses the movement from an understanding of the mechanisms claimed to be involved in intervention causation to the creation of a list of markers that evaluations need to report in order to facilitate an argument for the transferability of their results to any target context. Section Four discusses how that list of markers was used to generate the dataset of final results that is described and discussed in Chapters Seven and Nine.

6.1 IDENTIFYING THE SAMPLE OF EVALUATIONS

The second-most studied intervention-outcome pair identified by analysis of the AidGrade database was deworming for weight, with 18 evaluations identified. In accordance with the research protocol outlined in Chapter Four, this observation determined the selection of the set of evaluations of the effect of deworming on weight as the second case to be examined.

As with the first case examined, the second step was to search the 3ie repository of (quasi-)experimental impact evaluations to identify more recent, best-practice evaluations to add to the sample. Unlike for conditional cash transfers, there is no intervention ‘sector’ tag corresponding to deworming in the 3ie database. For this reason it was necessary to use the search function to identify database short entries related to deworming evaluations that could be screened for publication date before full texts were screened to ensure only inclusion of (quasi-)experimental impact evaluations of the effect of deworming interventions on child weight (or equivalent outcome measures such as BMI). In order to develop a suitable search string, first, the original AidGrade search protocol was consulted. On consulting this document it transpired that the original search string employed only contained the terms ‘deworming’ and ‘de-worming’ as well as other terms intended to restrict the results to impact evaluations employing an experimental or quasi-experimental design (AidGrade, 2013c). An initial search of the 3ie repository using just these terms returned 23 results. By interrogating these results, as well as the deworming evaluations in the AidGrade dataset, some more terms were identified that might plausibly increase the number of search results. Because of the maximum length of search strings accepted by the repository’s search tool, these terms could not all be included in one consolidated search. Rather than conducting multiple searches and having to resolve duplicates, each term was tested in turn, and only those terms that resulted in more results being returned were included. The full list of terms and a record of whether they were included in the search string or not is presented in Table 6.1. Any future search seeking to update this work should consider re-testing this list as well as attempting to identify more terms that may have entered the literature and may be useful in identifying studies.

Table 6.1: Search terms tested

Search term	**Included/excluded**
worm	included
helminthiasis	excluded
helminth	included
hookworm	excluded
roundworm	included
whipworm	excluded
nematode	excluded
nematodiasis	excluded
mebendazole	excluded
albendazole	included
ascariasis	excluded
ascaris	excluded
trichuris	excluded
trichuriasis	excluded
necator	excluded
necatoriasis	excluded
ancylostoma	excluded
ancylostomiasis	excluded

The longer search string thus developed resulted in an increase in results returned from 23 to 47 evaluations. As with the first case, these results were then screened by date published. Abstracts of studies published up to 2013 were screened to identify any evaluations not included in the AidGrade database but which might be relevant. This identified that Miguel and Kremer (2004) had been excluded from the AidGrade database, presumably because no treatment effect of helminth treatment alone was estimated; all treatment effects included treatment for schistosomiasis, also. However, as an effect of a treatment including deworming for helminths is estimated, reporting the contextual and intervention features of relevance to the functioning of the mechanisms involved is necessary. Given that, this evaluation was also a legitimate candidate for analysis as a member of the set and so was included, bringing the number of evaluations in the set to 19.

Sixteen evaluations were identified in the 3ie database published in 2013 or later and not included in the AidGrade database. The full texts of these 16 evaluations were then examined in order to include only evaluations where the intervention included mass deworming of children and the outcomes measured included weight or an equivalent measurement such as BMI. This full text screening resulted in four studies included, bringing the total number of evaluations in this set to 23. One paper listed in the AidGrade database could not be located and so was not included, bringing the total number of evaluations in the set to 22. Descriptive statistics on this set of evaluations are available in Table 6.2, below. A full list of the evaluations included in the set for case two is available in Appendix B.

Table 6.2: Descriptive statistics of evaluations for case study two

Total	**AidGrade**	**3ie**	**RCT**	**Other Method**	**Before 1980**	**1980-89**	**1990-99**	**2000-2009**	**After 2009**
22	17	5	22	0	0	2	7	8	5

6.2 IDENTIFYING THE INTERVENTION THEORY OR THEORIES

As with the first case study, the next step in examining this case was to identify the model or models of intervention causation that predominate in the literature in order to synthesise them and create a realist interpretation of the programme theory or theories behind the set of evaluations.

6.2.1 Purposive sampling of the relevant literature

As with the first case, this process began by extracting the material reported about theoretical mechanisms by each evaluation in the set. Every full text was examined and this material extracted. As with the first case study, 100% of included evaluations were examined in order to ensure that no theories pertaining to any evaluation were missed. Once this information had been gathered, systematic reviews including evaluations in the set and literature reviews referencing evaluations in the set were identified using web searches. As with the first case described in the previous chapter, no pretence of exhaustivity is implied here. Those reviews that were easily identifiable were consulted. Other reviews may have been missed. Exhaustivity was not necessary at this stage because the purpose of this stage in the literature review was merely to increase the efficiency of the search process, allowing the theory map to reach a point of practical adequacy more quickly than if those reviews were not consulted. The third and final stage in the literature review process was to consult each evaluation in the set in a randomly generated order, follow its theoretical references and their references until no new theoretical information was being added to the theory map in progress. Once the theoretical references of three evaluations in a row revealed no new information, the theory map in progress was judged to have reached a point of practical adequacy and the literature search was terminated.

This purposive literature sampling strategy led to the investigation of all 22 evaluations in the set, nine reviews, and 30 theoretical references from the reviews and from five root evaluations in the set before the theory map presented in Table 6.3 emerged and remained unchanged for long enough to be considered practically adequate.

6.2.2 Generating the theory map

As with the conditional cash transfer literature, the theoretical frameworks underpinning evaluations in the set were found to be remarkably homogeneous, facilitating the creation of a single theory map that could be used to assess the engagement with and reporting of context for every evaluation in the set. However, one marked contrast with the case of conditional cash transfers for school enrolment is immediately apparent. In the case of conditional cash transfers, all the mechanisms identified in Table 6.3 act at a single point in time, the point at which a decision is made to enrol a child in school or not. At this point in time, a single event occurs, which occurs differently because of the action of the mechanism identified. In this way, these mechanisms are composed of single causal links that connect the context at one moment in time, as modified by the intervention, with an outcome shortly after. By contrast, the mechanisms identified in Table 6.3 are more extended in time, and are composed themselves of chains of events, connected by multiple causal links that transmit the influence of the mechanism over several sequential pairings of context and intermediate outcome.

It might be objected that the mechanisms identified in Table 6.3 are 'not really mechanisms' because they are composed of causal chains that themselves seem to contain mechanisms. However, I do not believe that this sort of reductionism is warranted. Consider some canonical examples of mechanisms from the realist literature. Pawson and Tilly (1997) famously consider the case of CCTV in car parks intended to reduce the incidence of car crime. This intervention is said to plausibly lead to changes in the outcome through at least eight mechanisms, which Pawson and Tilly (*ibid*, p.78) list and describe. One is described as follows:

> *(c)* The 'nosy parker' mechanism. *The presence of CCTV may lead to increases in usage of car parks, because drivers feel less at risk of victimization. Increased usage could then enhance natural surveillance which may deter potential offenders, who feel they are at increased risk of apprehension in the course of criminal behaviour.*
>
> (*ibid*)

This description, like the mechanisms described in Table 6.3, describes a chain of causal reasoning. The intervention, CCTV, combines with the context to set off a cascade of causal effects that includes effects on the outcome amongst other things. Pawson and Tilly's mechanism, then, describes a pathway through which the intervention might affect the outcome, but can be unpacked into a chain of lower-level effects. These lower-level effects are themselves connections between contexts and outcomes that rely on the causal powers and liabilities of these lower-level contexts as described. So, for example, the first link in the chain of causation described by Pawson and Tilly (*ibid*) as the ‘nosy parker’ mechanism is that ‘drivers [become aware of the CCTV and] feel less at risk of victimisation’. This relationship between drivers’ awareness of CCTV and their feeling of risk itself requires explanation and can be explained by a lower-order mechanism we could call the ‘reassuring presence of authority’ mechanism, or any one of several other names. It may be useful, even necessary, to drill down into such lower order mechanisms in order to uncover essential features of the context that enable the action of the lower-order mechanism and therefore are also essential for the operation of the whole chain that constitutes the higher-order mechanism. For example, in this case, an expectation among car-park users that the sorts of authorities likely to be monitoring the CCTV are effective and interested in deterring crime is necessary for the ‘reassuring presence of authority’ mechanism to activate. This in turn is necessary for the ‘nosy parker’ mechanism to activate. In some contexts this will be more true than in others. The decomposition of higher-order mechanisms into chains of lower-order CMOs also allows us to identify intermediate outcomes like a reduced feeling of risk among carpark users. Interrogating and reporting such intermediate outcomes allows us to build compelling causal explanations by confirming or eliminating possible mechanisms through which causation might be operating. Intermediate outcomes that might be usefully reported for evaluations of deworming for weight are discussed in Subsection 6.3.2.

So, it is useful for the purposes of understanding CCTV in carparks to reduce car crime or deworming of children to increase weight to talk about higher-order mechanisms that connect the intervention, in combination with context, to the outcome. To understand the full set of contextual features that are relevant to intervention causation, it is also useful to attempt to unpack those mechanisms into more complex causal chains. This is because it allows us to examine the lower-order mechanisms that are necessary for the higher-order mechanism to function and allows us to identify barriers or enablers to the action of the intervention that might not be obvious at the higher level. The mechanisms identified in Table 6.3, then, are indeed mechanisms in the realist sense, despite the fact that it does appear legitimate to also apply the term ‘mechanism’ to the processes taking place at each point along the causal chain of which they are constituted. Though Pawson and Tilly do not talk about such ‘nesting’ of mechanisms, some of their mechanisms take this form. Others don’t and are more like the mechanisms described for the previous case in Table 5.2.

The theory map as presented in Table 6.3 emerged very quickly from the consultation of the evaluations in the set and remained unchanged throughout the consultation of the rest of the evaluations and the consultation of theoretical references. More detail was still being added at a more precise level, however. In particular, elaborating on the lower-order mechanisms that lead from decreased worm burden to increased weight for an individual required interrogating a lot of sources to dig into a level of detail that was entirely absent from the evaluations and from most theoretical references. This was necessary in order to identify all of the markers of context that are relevant to intervention causation.

The complete table of CIMOs is presented as Table 6.3 below, with each element from the table explained in the following section. Each CIMO can be read across one row of the table. As the contextual and intervention features apply to both mechanisms, these columns span both rows.

Table 6.3: Final programme theory map represented as CIMOs

Contextual feature	**Intervention feature**	**Mechanism**	**Outcome**
Children with some level of malnourishment leading to their being under-weight.[44] AND Significant levels (in prevalence and/or intensity) of soil-transmitted parasitic worm infection (helminthiasis) by the species *Ascaris lumbricoides* (roundworm), *Trichuris trichiura* (whipworm), *Necator americanus* or *Ancylostoma duodenale* (both types of hookworm).[45]	Administration of an anthelmintic (deworming) drug, most commonly albendazole or mebendazole, either targeted to infected children, to an at-risk population, or to the whole population through mass drug administration (MDA) to all children.[46][47][48]	**Direct chemotherapy** Degenerative alterations in the tissues of the worm, leading to worm immobilisation and death, leading to decreased worm burden, leading to increased nutrient uptake and retention.[49]	Increased weight of dewormed children.
		Reduced reinfection Decreased worm burden in the treated (by the process above), leading to fewer eggs in faeces, leading to fewer parasites in soil, leading to a lower rate of reinfection, leading to increased nutrient uptake and retention.[50]	Increased weight of all children.

[44] See, for example, Hall (1993) on the insufficiency of simple worm infection for a negative effect on weight and other developmental indicators.

[45] See, for example, Zhang *et al.* (2017) and Majid *et al.* (2019). Hall (1993) is illuminating on the importance of intensity. Subsection 2.3.1 explains the exclusion of schistosomiasis from the theory map.

[46] See, for example, Gabrielli *et al.* (2011).

[47] Older evaluations sometimes use older anthelmintic such as piperazine.

[48] This latter is sometimes referred to in the literature as preventative chemotherapy (PC), but the term MDA is more popular and equivalent.

[49] See, for example, Stephenson *et al.* (2000), especially p.S27.

[50] See, for example, Dossa *et al.* (2001), Zhang *et al.* (2017)

6.2.3 Presenting the elements of the theory map.

This section provides more detail on the CIMOs described briefly in Table 6.3. First, however, elements of theory that were considered and excluded are discussed.

6.2.3.1 Elements considered and excluded

Discussion of the theory map for the case of CCTs for school enrolment included lengthy consideration of theoretical elements of the literature that were not relevant for inclusion in the theory map generated for this research project. This was partly necessary because the theoretical literature relating to CCTs is extensive and accords a lot of importance to topics that were not relevant for present purposes. Examples include the important discussions in the literature of the political economic considerations that might justify an unconditional cash transfer (UCT) over a CCT, or the downstream effects of increasing school enrolment. By contrast, the theoretical literature addressing deworming is much less extensive and much more focussed on questions of intervention effectiveness. This means that much less discussion is needed in this section than was needed in the equivalent section of the previous chapter. Nevertheless, some important theoretical discussions in the literature were not relevant for present purposes, and those are discussed in this subsection.

6.2.3.1.1 Schistosomiasis treatment

As discussed, the AidGrade database v1.3 was used to identify the pairing of ‘deworming’ and ‘weight’ as a well-studied intervention-outcome pair. The intended referent of the term ‘deworming’ is not immediately clear from the AidGrade database, the accompanying technical documents, nor the Vivalt (2019) paper. Some authors use the term ‘deworming’ to identify not just treatment of soil-transmitted helminths (STH) such as roundworm, whipworm and hookworm, but also water-transmitted flatworms also known as schistosoma which give rise to a disease known as schistosomiasis, bilharzia or snail fever. However, schistosomiasis is a quite different disease which operates and manifests in a different way to helminthiasis and is treated differently (Welch et al., 2017). Although this treatment is sometimes conducted in parallel with treatment for helminthiasis, the mechanisms of action of both treatments are quite distinct, employing different targeting strategies, different drugs, and targeting different outcomes. Given these facts, it is unsurprising that full-text consultation of the evaluations identified in the AidGrade database revealed those evaluations to be concerned only with treatment of helminthiasis rather than schistosomiasis. When searching the 3ie (quasi-)experimental impact evaluation repository to add more recent studies to the set, this example was followed and only evaluations of interventions targeting soil-transmitted helminth infections were included.

6.2.3.1.2 Other outcomes related to increased nutritional status

Much of the literature related to the deworming treatment of helminthiasis-afflicted populations is concerned with a broad range of outcomes of which weight is only one. In particular, many evaluations attempt to identify effects of deworming treatment on increased cognitive function, school attendance, and even lifetime earnings.[51] Like weight, these outcomes are downstream effects of the improved nutrition of treated children. However, because they are further downstream than weight, with many more links in the causal chains that connect them to deworming, their causation is much more complicated, being exposed to more barriers and enablers at each of the links in these long causal chains. In identifying the evaluations to be included in the set, and in developing the theory map, the analysis here was limited to the intervention-outcome pairing of deworming and weight.

6.2.3.1.3 Other accompanying interventions

Just as care was necessary to restrict the theory map to only theory of relevance to the causation of one outcome, weight, it was necessary to restrict consideration to one tightly defined intervention, STH-targeted deworming of children. Many of the evaluations in the set were evaluating deworming in combination with supplementation, for example of iron or vitamin A. Theory relating to the effect of these accompanying interventions on outcomes was not relevant to the ultimate project of this strand of the current research project, building a map of the theoretical understanding(s) of the causal effect of deworming on child weight shared by

[51] See, for example, Miguel and Kremer's seminal paper (2004) and the insightful overview of the literature provided by Majid *et al.* (2019).

evaluations in the set in order to assess the engagement with and reporting on context of those evaluations. Therefore, this theory was not included in the programme theory map represented in Table 6.3.

6.2.3.2 CIMOs organised by mechanism

This subsection provides a more detailed account of the two CIMOs summarised in Table 6.3 in order to facilitate an argument for the identification of all and only those contextual features and intervention features that will be identified in the following section as necessary for an argument for the transferability of evaluation findings.

6.2.3.2.1 Direct chemotherapy

The principal mechanism by which deworming affects weight has been labelled ‘direct chemotherapy’ in Table 6.3. This term was chosen so as to encompass the different treatment targeting strategies which are represented in the evaluations in the set. In one approach, only infected children are targeted for deworming, whereas in another, an at risk group is targeted. In the most common approach, all children in the population are treated. This latter strategy is referred to as the preventative chemotherapy (PC) approach, or mass drug administration (MDA). As MDA is the more frequently used term, it is used in this thesis.[52] Because the anthelmintic drugs used in deworming are very low-cost and low-risk, but diagnosis of worm infections is expensive, MDA is considered a lower-cost equivalent to treating only infected children (Silva et al., 2015).

Whichever of these treatment-targeting approaches is used, the mechanisms of action of the treatment are the same. This follows because in either case, deworming treatment only affects the infected, and does so in the same way. This is important to establish, because if it were not the case the programme theories would not be compatible between the different targeting

[52] The claim about frequency of use follows Welch *et al.* (2017), who conducted a more exhaustive review of the literature than that carried out for this research project.

methods, and it would be necessary to assess the reporting on context of each subset of evaluations separately.

The association between infection with soil transmitted helminths and lower nutritional status was made in the early 20th century (Smillie and Spencer, 1926; Smillie and Augustine, 1926). In time, the biological theory explaining the mechanisms through which worms caused lower nutritional status was elaborated and is now well understood. Stephenson *et al.* (2000) review this theory and summarise the most important mechanisms as direct nutrient loss e.g. due to fecal or urinary blood loss, decreased appetite due to digestive discomfort and malfunction as well as misguided attempts to treat these through starvation, and decreased nutrient uptake and utilisation e.g. fat malabsorption due to mucosal damage. Removing these negative effects by clearing the intestines of worms should lead to higher nutritional status and increased weight amongst underweight individuals. These effects will be most marked in children during the crucial first 24 months of life. Analysis of worldwide trends in growth faltering shows that weight and height can be influenced much more successfully in these first two years than at any other time (Victora et al., 2010).

The prevalence of worm infections as well as the intensity of those infections will determine the expected magnitude of effects of deworming on weight. Only moderate and high intensity infections have been shown to cause adverse effects on nutritional status generally and weight specifically (Majid, Kang and Hotez, 2019). Attempts to increase weight will also be facilitated by the level of children's growth deficit, the quality of their diet, and the extent to which they are free from other disease (Dossa et al., 2001; Sarkar et al., 2002).

Which anthelmintic drug is administered and in what dose is also relevant to the operation of the mechanism, especially in combination with the level of infection of the different species of parasitic worm in the population. Reviews of efficacy evaluations of different deworming drugs focussed on level and rate of worm infection as their outcome have shown that the different drugs available are differently effective against different species of worm. For example, albendazole and mebendazole are both highly effective against *ascaris lumbricoides* and ineffective against *trichuris trichiura*. Mebendazole is ineffective against the hookworms, whereas albendazole can be effective, albeit less so than against *ascaris lumbricoides* (Keiser and Utzinger, 2008). For this reason, the infection prevalence and intensity of each separate species of worm will be relevant to the extent to which the intervention should be expected to affect the outcome.

6.2.3.2.2 Reduced reinfection

The second mechanism through which deworming can increase the weight of children branches off from the action of the first mechanism at the point of reduced worm burden in treated individuals. While this has a direct effect on nutritional status as described in the previous subsection, it also has an indirect effect by way of a different causal chain. This indirect effect occurs because constant reinfection with worms is important to the effect of STH infection on weight. In contexts where the infected are being treated, they will excrete fewer worm eggs. As these eggs are an important vector of (re)infection, this will indirectly reduce the burden of disease in the whole population (Zhang et al., 2017). The dynamics of reinfection are why deworming programs generally treat participants repeatedly, classically every 6 months, but sometimes more or less frequently (Gabrielli et al., 2011).

Transmission of parasites from fecal matter to new hosts occurs in the context of open defecation. The eggs in faeces then pass into the soil and are either ingested directly (through children playing in soil and transmitting eggs from hands to mouth, or on unwashed vegetables, or through water sources) or stay in the soil long enough to hatch, at which point larval worms are capable of infecting children through their bare feet. The nature of sanitation facilities, the footwear of children, the type of water sources used, and hygiene practices are therefore all relevant to the extent to which this mechanism should be expected to function (Gabrielli et al., 2011). In the context of widespread latrine use, good hygiene, safe water source usage, and/or use of footwear by children, the rate of reinfection should be much lower, and reducing this rate much less effective (Gabrielli et al., 2011; Zhang et al., 2017).

Also important to the action of both mechanisms is the competence of the implementing agency, especially when measuring effects using an intention-to-treat approach rather than one which looks at the treatment effect on the treated. Attrition between baseline and successive rounds of follow up is a natural feature of public health intervention, and with a periodic, repeated treatment like deworming will have a large effect on effectiveness at the population level (Awasthi et al., 2013; Ebenezer et al., 2013).

6.3 FROM PROGRAMME THEORY TO DETERMINANTS OF TRANSFERABILITY

The previous section identified the principal mechanisms responsible for the change in child weight due to deworming. At this point, it is worth reiterating that the ultimate goal of this aspect of this research project is to use this case study to examine the systemic differences, if any, between (quasi-)experimental impact evaluation methods regarding the extent to which they facilitate the transferability of their findings to some other context. Chapter Three argued for the validity and the utility of the tools of realism to answer this question and gave it a theoretically rich interpretation:

> *What are the systematic differences, if any, between (quasi-)experimental impact evaluation methods regarding the extent to which they explore and report on the barriers and enablers of intervention mechanisms present in the study context and the extent to which different intervention mechanisms have been activated?*

The way in which this has been investigated is through the examination of two sets of evaluations employing diverse (quasi-)experimental impact evaluation methods. As Table 6.2 shows, this second case study does not contain the methodological diversity of the first. Therefore the insights arising from this case must be combined with those arising from the first case in order to inform an answer to research subquestion one. That combination is achieved in Chapter Nine. In order to render this case study informative for the attempt to answer research subquestion one, each evaluation in the set needed to be assessed on the extent to which it reported on 'the barriers and enablers of intervention mechanisms present in the study context and the extent to which different intervention mechanisms have been activated.'

The next step in the analysis of this case, then, was to move from the programme theory map given in Section 6.2 to an enumeration of the information that would have to be reported by evaluations of the effect of deworming on weight in order to facilitate an assessment of 'the barriers and enablers of intervention mechanisms present in the study context and the extent to which different intervention mechanisms have been activated.' This was a two-stage process that began with an assessment of the barriers and enablers of intervention mechanisms and concluded with an assessment of the information required to determine the extent to which different intervention mechanisms had been activated.

6.3.1 Barriers and enablers of intervention mechanisms

Reflecting on the programme theory map as outlined in Table 6.3 and elaborated upon in Section 6.2.3.2 makes clear that the barriers and enablers of intervention mechanisms can helpfully be divided into features of context and features of the intervention as implemented. These features are depicted in Table 6.4, organised by mechanism. Barriers or enablers that apply to both mechanisms are shown in columns that spread across both rows in the table. Barriers or enablers that apply only to one mechanism are shown in a cell that only extends one row by one column.

Table 6.4: Barriers to and enablers of the operation of intervention mechanisms

Mechanism	**Contextual barriers/enablers**		**Intervention barriers/enablers**
Direct chemotherapy		Baseline nutritional status of children General burden of disease in child population Infection prevalence and intensity by worm species Quality of diet of children Age of children	Drug administered Dose Frequency Targeting of treatment[53] Implementation agency
Reduced reinfection	Hygiene practices Sanitation facilities Water sources Child footwear		

6.3.2 Intermediate outcomes as evidence of the action of mechanisms

In the previous chapter, discussing the previous case, it was argued that disaggregated reporting of the outcome of interest would facilitate distinguishing between the action of the different candidate mechanisms for the effect of the intervention on the outcome. As discussed in Section 6.2.2, the mechanisms active for that case are different to those active in this case by all acting at a single point in time on a single link in a causal chain. By contrast, the mechanisms described for this case can be disaggregated into multi-link causal chains. One consequence of that difference in the structure of the causal model is that, for this case, various intermediate outcomes manifest themselves at different points on those causal chains. For example, while child weight is the final outcome of interest, prevalence and intensity of worm infection are both intermediate outcomes that should be altered by the intervention. Examining and reporting on

[53] Whether the intervention has been targeted to only children diagnosed with an infection, to an at-risk population, or to the whole population is clearly relevant to the magnitude of the expected effect. Much of the controversy in meta-analysis of these evaluations is that absence of an average effect over a large population is not evidence of a null effect for infected individuals, but it has sometimes been described in that way (Silva et al., 2015).

these intermediate outcomes will be essential for the facilitation of any argument for the transferability of findings of an evaluation in this set to some other context. This is because changes in intermediate outcomes can tell us how the intervention worked or did not work and why. For example, consider a case in which administration of anthelmintic drugs led to no detectable change in child weight. Was this perhaps because drug distribution only reached a limited number of intended beneficiaries? Or was it perhaps because reinfection rates were too high for periodic deworming to have any effect? If proportion of intended beneficiaries who received medication in each round is reported (intermediate outcome 1), and prevalence and intensity of infection at endline or even at every round of treatment are reported (intermediate outcome 2), it may be possible to build an argument for one of these competing explanations. For this reason, evaluations in the set are assessed on the reporting of these two intermediate outcomes as well as the contextual features and intervention features that are barriers or enablers of the intervention's effect on the outcome.

The research protocol outlined in Chapter Four calls for the qualitative and quantitative comparison of evaluations in the case study set regarding the extent to which they 'explore and report on the barriers and enablers of intervention mechanisms present in the study context and the extent to which different intervention mechanisms have been activated.' This section has provided a list of the barriers and enablers of intervention mechanisms as well as two intermediate outcomes which should be reported in order to facilitate an argument for the extent to which changes in outcome variable can be attributed to the action of each of the two mechanisms identified. Armed with this information, the evaluations in the set could be analysed qualitatively and quantitatively on the nature and extent of their reporting of these barriers, enablers and outcome measures.

6.4 GENERATING A QUANTITATIVE AND QUALITATIVE ASSESSMENT OF THE FACILITATION OF TRANSFERABILITY

The previous section has presented an argument for the existence of a set of barriers and enablers of intervention causation as well as two intermediate outcomes which should be reported by an evaluation in the set in order to facilitate an argument for the transferability of results to some other context. These barriers, enablers and axes of disaggregation can be simplified to generate a list of causal markers, as in the previous case. The reporting of these markers can then be assessed for each evaluation in the set in order to facilitate the comparison of evaluations which is necessary to answer the research question addressed by this strand of research. The full list of markers against which evaluations in the set were assessed is reproduced below for clarity:

Table 6.5: Deworming MICCs

Group	**Subgroup**	**Marker**
Context	Environment context	Hygeine practices
		Sanitation facilities
		Child footwear
		Water sources
	Biological context	Diet quality
		Burden of other disease
	Baseline non-worm	Baseline nutritional status
		Baseline age
	Baseline worm	Baseline infection prevalence
		Baseline infection prevalence by species
		Baseline infection intensity
		Baseline infection intensity by species

Intervention	Drug administered
	Dose
	Frequency
	Targeting (MDA/at risk/diagnosed)
	Implementation agency
Intermediate Outcomes	Proportion treated
	Endline prevalence
	Endline prevalence by species
	Endline intensity
	Endline intensity by species

Each evaluation in the set was examined and scored 1 if it reported a given marker, 0 if it did not. In addition to the score variable, qualitative assessments of the reporting were also collected as a separate variable for each marker. In addition to this information, the ability of an evaluation team to generate data was also assessed and recorded as either medium, high or very high along with a qualitative explanation of why the score had been given. The use of three categories for ability to generate data was established in piloting as a satisfactory modelling simplification with adequate power to express differences between the evaluations in the set. ‘Medium’ was recorded for evaluations that caused non-standard anthropometric measurements to be generated for children, with or without stool sample analysis, but did not conduct surveys of households or children. The term ‘Low’ was not chosen to represent this category in order to increase comparability with case one, that of CCTs for school enrolment. In that case, ‘Low’ was used for evaluations that relied on the reinterpretation of secondary data, data that the evaluators had not caused to be collected. None of the evaluations in case two fall into this category, as all caused some additional data to be generated. ‘Medium’ was therefore used to label the category in which anthropometric measurements were caused to be collected, and ‘High’ was used for evaluations that additionally conducted surveys. Most evaluations fell into one of these two categories, with nine designated ‘High’ and ten designated ‘Medium’. ‘Very high’ was recorded for three evaluations that used more costly survey practices to engage deeply with context. In one case (Zhang et al., 2017) this was a focus-grouping exercise with study villagers to generate qualitative information about health practices. In the other two cases, (Donnen et al., 1998; Joseph et al., 2015) these were household surveys conducted at the household’s place of residence. This method of assessing the evaluations in the set resulted in a dataset with 22 observations and 46 variables, creating 1,012 data entries. The full dataset is available at https://mattjudendotcom.files.wordpress.com/2021/04/mjuden_thesis_data_and_code.zip and has been uploaded to the UK Data Service where it has been deposited in accordance with the conditions of my ESRC studentship under Project ID 204604.The final results of the quantitative and qualitative assessment of evaluations in the set are reported in detail and discussed in Chapters Seven and Nine.

7 The method generates informative insights for both cases and is therefore successful

The previous two chapters have described in detail how two datasets were created, one for each of two cases, where each case corresponds to a set of evaluations of a well-studied pairing of intervention and outcome. These datasets were created in order to answer the operationalised form of research subquestion one which was developed in Chapter Four, Subsection 1.3. The operationalised forms of both research questions are recapped in the box below for convenience.

Research question recap

Research question	**Answered in**
Research subquestions:	
1.	
a) Can realist programme theory mapping be adapted to create a tool to assess the transferability of (quasi-)experimental development impact evaluation results?	Chapter 7
b) If so, what can it tell us about the systematic differences, if any, between (quasi-)experimental impact evaluation methods regarding the extent to which they report on the barriers and enablers of intervention mechanisms present in the study context and the extent to which they report the degree to which different mechanisms are responsible for changes in outcomes…	
i. as they are currently used? and	Chapter 9
ii. as they might be used?	Chapter 9
2. For epistemic communities of development experts, what are the shared notions of validity concerning what counts as a 'high quality' (quasi-)experimental impact evaluation? Further, what are the features of these accounts that are valued by members of the community, and what unresolved puzzles or nascent crises undermine them?	Chapter 8
Primary research question:	
Can we give a useful, systematic account of the relative merits of evidence generated using different (quasi-)experimental development impact evaluation methods that goes beyond internal validity to also consider the extent to which methods facilitate the transfer of results to other contexts? If so, how?	Chapter 9

In order to answer 1. a) it is sufficient to demonstrate that the data generated by applying the method developed in Chapter Four to each of the two cases does indeed permit an informative assessment of the transferability of results for evaluations in the two sets. This chapter provides such a demonstration. In the first and second section, for each case respectively, the method is demonstrated to generate useful insights for evaluators, policy makers and those attempting to synthesise the available evidence on the impact of interventions. Some examples are given of how small design changes for evaluations in the set could have rendered them much more powerfully able to engage with and report on markers of intervention causation in context (MICCs) so as to facilitate the transferability of their findings to other contexts. In Section Three, limitations of the analysis of Case One and Case Two are discussed, such as the fact that many of the evaluations included in each set were designed for purposes quite different to contributing to middle-range theory about how the intervention works. Section Four concludes that research subquestion 1. a) can be answered in the affirmative. That is, realist programme theory mapping has been adapted to create a tool to assess the transferability of (quasi-)experimental development impact evaluation results.

7.1 WHAT HAS BEEN LEARNED ABOUT EVALUATIONS IN CASE ONE?

This section discusses the marker scores generated for evaluations in Case One, evaluations of conditional cash transfers for school enrolment. This is achieved through the quantitative interpretation of the dataset of marker scores constructed as described in Chapter Five, as well as enriching this understanding through an incorporation of the qualitative information generated. In assessing the success of the method, it is helpful to begin at the most basic level. As this section progresses, the analysis becomes progressively more fine-grained. Therefore, the distribution of total marker scores over all evaluations is assessed first. Second, the distribution of total scores is assessed as disaggregated by method and by year. Next, the distribution of scores for each group of markers and then for each subgroup of markers is assessed. Lastly, the reporting of individual markers by the evaluations in the set is considered, and then the reporting of all markers by particular evaluations.

Before beginning the analysis of this section it is helpful to remember what the markers are for Case One, and how these markers are grouped into groups and subgroups. Therefore, Table 5.4 from Chapter Five is reproduced below. In that table we can see the full list of markers as well as the group and subgroup divisions.

Table 5.4: CCT MICCs

Group	**Subgroup**	**Marker**
Context	Level of enrolment	Household perceived privately optimal level of enrolment
		Enrolment preferences and/or rationality measures disaggregated by HH member
		Baseline enrolment
	Financial barriers	Household available resources
		Direct costs of education
		Indirect costs of education
	Non-financial barriers	Erroneously low estimates of expected returns
		Failures of rationality - excessive future discounting etc.
		Familial or community norms
Intervention	Conditionality	Announcement of conditions
		Level of monitoring
		Enforcement of sanctions
	Transfer	Transfer recipient
		Size of transfer
		Regularity of transfer
		Implementing institution(s)
		Targeting criteria and method
Outcomes		Enrolment by HHH gender
		Enrolment by HH wealth/consumption
		Enrolment by marginality of child

7.1.1 There is considerable variety in the scores generated for evaluations

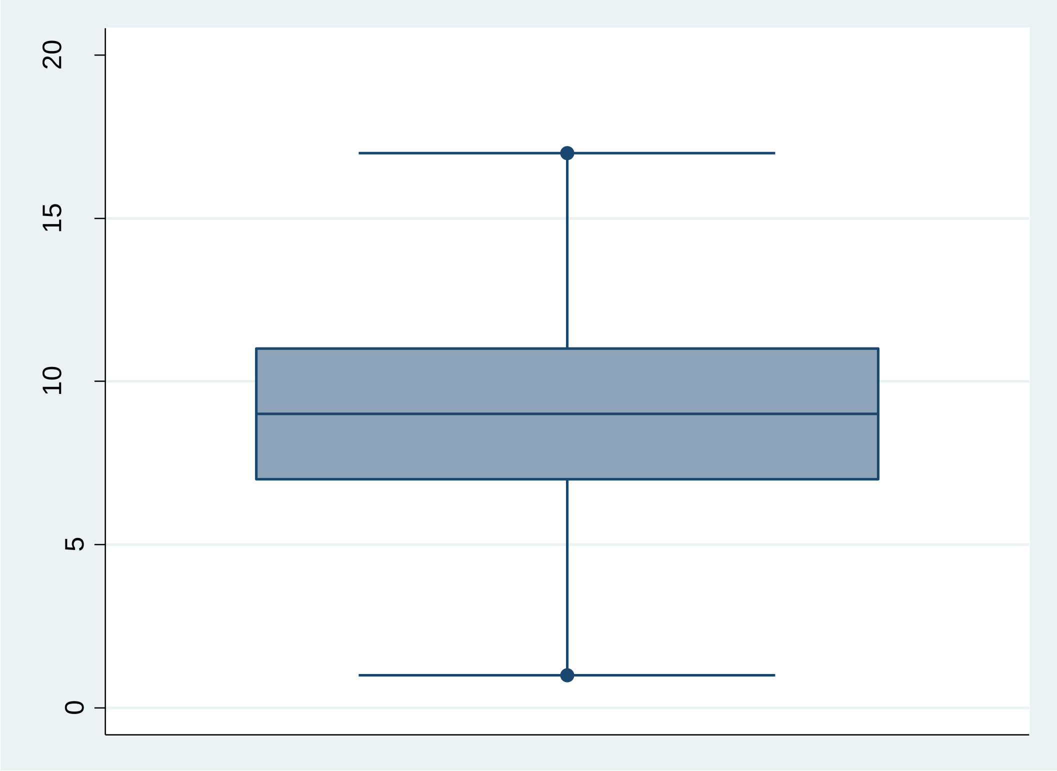
In judging the success of the method, it is informative to begin with an assessment of the total marker scores generated for evaluations in the set. If these scores were all low, it would be clear that the method had not generated a list of markers of transferability that evaluators in the set were ever motivated and able to report. As evaluators in the set are sensitive to the need to generate transferable insights, and as some of them are also convinced of the power of theory-based evaluation to provide a means to do this, it would be a worrying sign if none of them had reported many of the markers of transferability identified. This might indicate that the list of markers was somehow mis-specified, either by not in fact being entailed by the theory underpinning evaluations in the set, or by not being reportable in practice.

Similarly, if the total marker scores generated for evaluations in the set were all high, this might be a sign that the problem of transferability for evaluations in the set, and therefore possibly for (quasi-)experimental impact evaluations of development interventions in general, was in fact well addressed by current understandings and current tools, and the motivation for this research project outlined in Chapter Two was misinformed.

If total marker scores showed little variation, this might also be a worrying sign that the method was not very informative. This would suggest that, at least at the level of total score, the method was not able to distinguish examples of strong methodology that might be emulated and of weak methodology that might be avoided. While analysis of group and subgroup scores might have revealed variation, this would be a surprising finding if there was very little variation at the total score level.

It is therefore indicative of the success of the method that total marker scores exhibit a wide range and significant variation. The median total marker score for evaluations in the set is nine, with the highest score at 17 and the lowest at one. The first quartile of scores is at seven, and the third at eleven, giving an inter-quartile range of four. The median and inter-quartile range of scores is reported rather than the mean and standard deviation as the total score is a discrete, ordinal variable this is therefore more appropriate (Stevens, 1946, p.678). Figure 7.1, below, displays this information as a box plot.

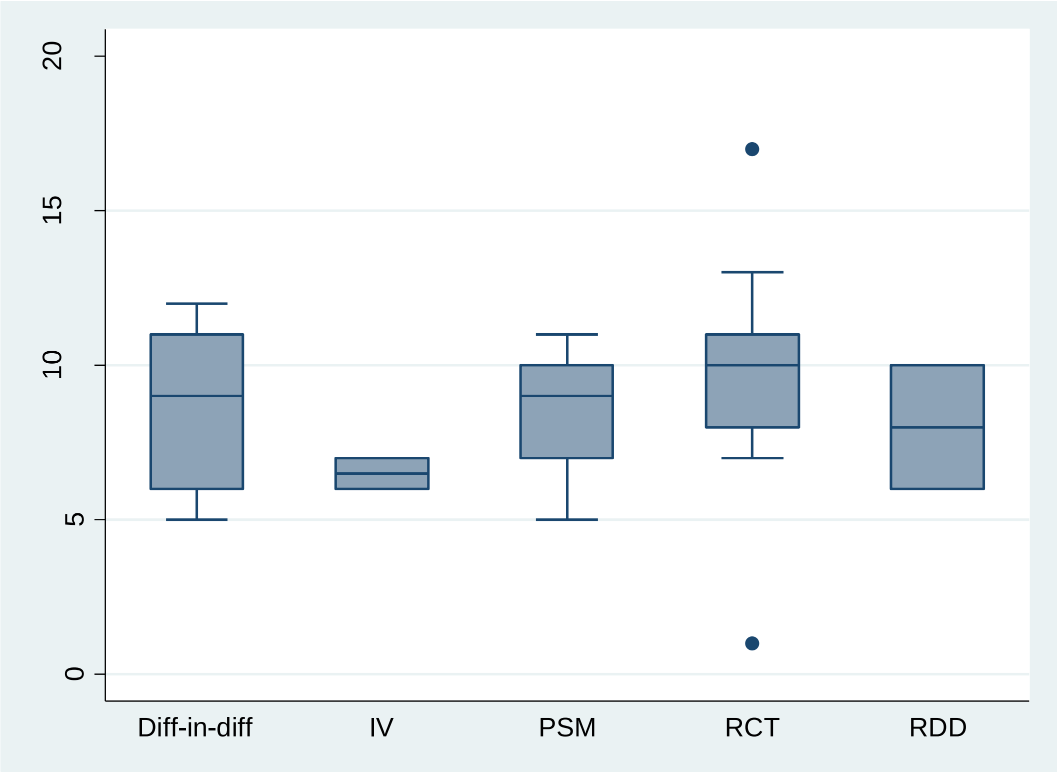
Figure 7.1: Distribution of total marker scores for evaluations in Case One



7.1.2 Total score is not associated with method choice

Another way in which the method developed in Chapter Four can be shown to be successful is that it can be shown to have produced data that is informative on the relationship between transferability and method choice. This relationship is informative to debates about the existence of a supposed trade-off between internal validity and external validity, as Chapter Nine will discuss in detail. Figure 7.2, below, shows the distribution of total marker shores for the evaluations in Case One disaggregated by method employed.

Figure 7.2: Distribution of total marker scores disaggregated by method for evaluations in Case One



It is clear from a visual inspection of the median values, variation and range of each subset of evaluations issued from a different method, that method choice does not seem to be driving total score. Although inter-quartile range is similar for Diff-in-Diff, PSM and RCTs, the RCT group demonstrates the largest range of values if outside observations are considered, as they should be. Note that outside values are not the same as outliers. The standard method I have used to calculate the inter-quartile range defines outside values, following Tukey (1977), as values greater than or equal to 1.5*IQR above the upper quartile or below the lower quartile. Especially with relatively few observations, there is no reason to exclude such values from the analysis. Although the range is largest for RCTs, this may partly be driven by the fact that there are more RCT observations in the set than for other methods, as shown in Table 7.1, below.

Table 7.1: Number of Case One evaluations employing each method.

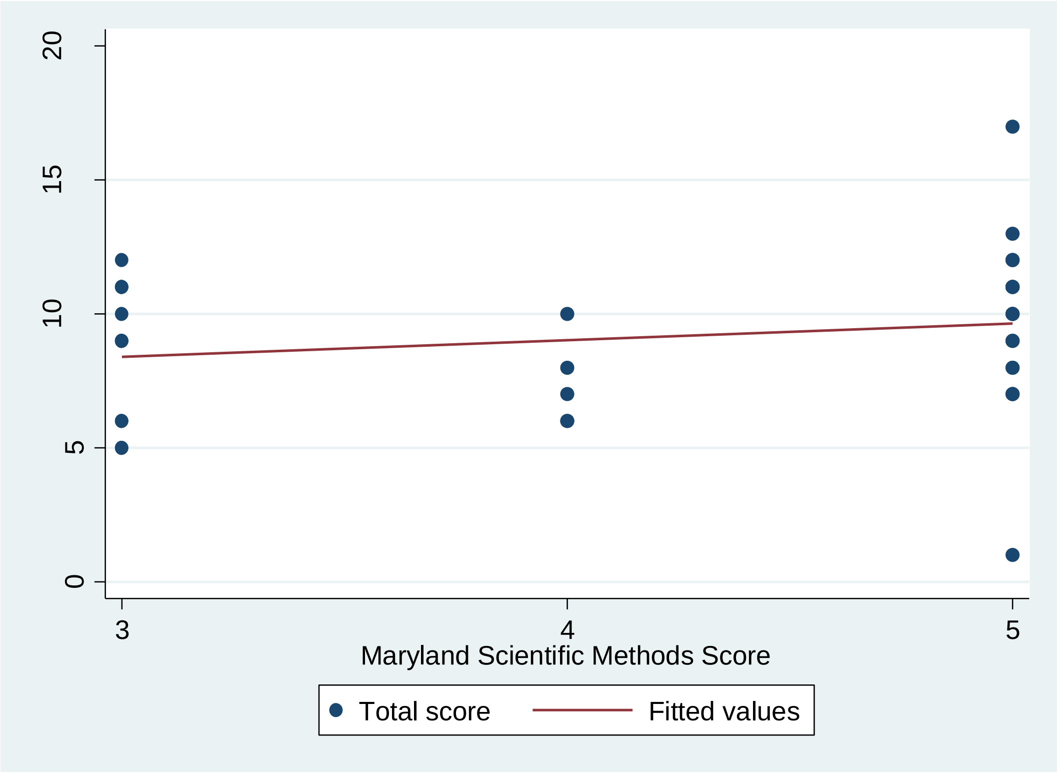
Total Evaluations	**RCT**	**Diff-in-diff**[54]	**PSM**[55]	**RDD**[56]	**IV**[57]
37	22	9	4	3	2

N.B. The sum of the methods counts is 40, reflecting the fact that three evaluations in the set employed two methods.

Figure 7.2 represents a treatment of the 'method' variable as a nominal variable, meaning as a variable that has no inherent order. However, as the discussion at the beginning of this section makes clear, we might consider the method variable to have an inherent order based on the level of facilitation of internal validity generally thought to be afforded by that choice of method. Though such an ordering is a simplification and often employed for illegitimate purposes, it nonetheless reflects a pervasive perception that some methods better facilitate internal validity than others. Though this perception is exaggerated, as Deaton and Cartwright (2018b) are foremost in arguing, it is not entirely incorrect. One tool that can be used to create such an ordering is the Maryland Scientific Methods Scale (MSMS) (Farrington et al., 2002). On this scale, RCTs are accorded a score of five, regression discontinuity designs (RDDs) and instrumental variables approaches (IVs) are accorded a score of four, difference in difference approaches (Diff-in-diffs) and propensity score matching approaches (PSMs) are accorded three. Creating a variable corresponding to the MSMS score as a proxy for internal validity allows the correlation between total marker score and internal validity to be estimated for evaluations in the set. Figure 7.3, below, depicts the association as a scatter plot with line of fit.

[54] Difference-in-differences approach
[55] Propensity score matching approach
[56] Regression discontinuity design
[57] Instrumental variable approach

Figure 7.3: Association between the Maryland Scientific Methods Scale and total marker score for evaluations in Case One



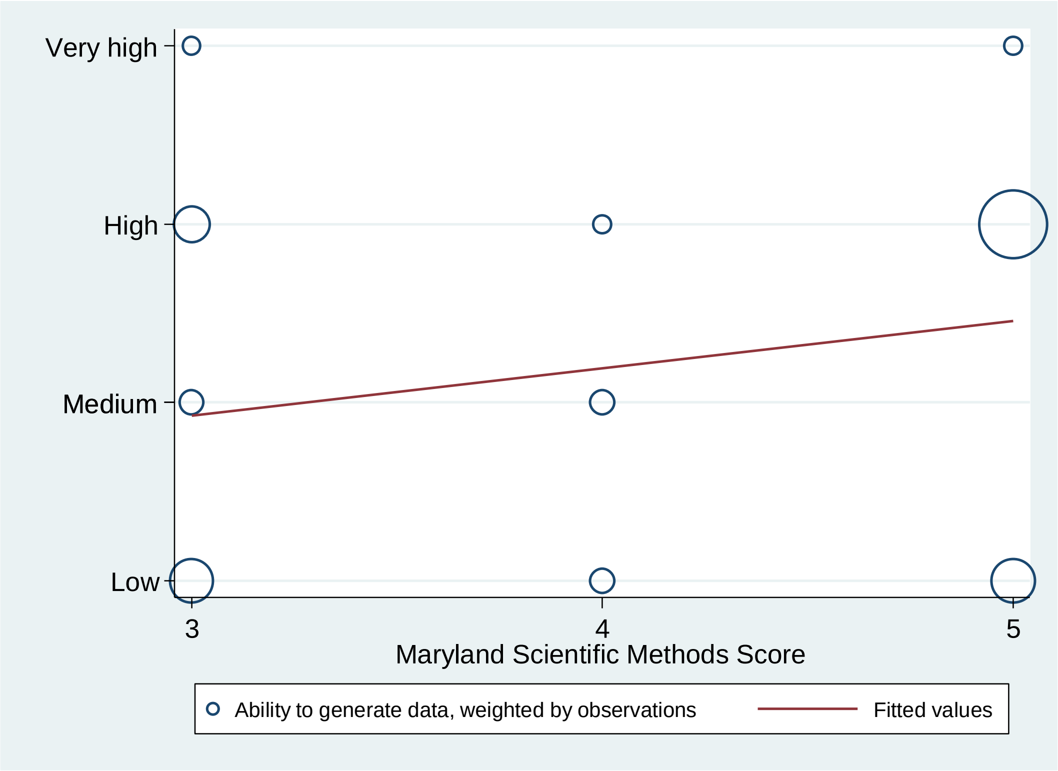
Calculating Spearman's Rank-Order Correlation results in a coefficient of 0.2552 with a p-value of 0.1120.[58] The correlation expressed by the coefficient is weak.[59] More importantly, the p-value indicates that there is an over 11% chance that such a result would occur by chance if the hypothesis of no association between the variables were true. Although any given threshold for statistical significance is arbitrary and should not be accorded as much importance as such thresholds generally are, an 11% chance of the result obtaining conditional on the truth of the null hypothesis suggests that the data are too few with too high a variance to reliably establish the presence of any correlation (McShane et al., 2019). However, by investigating the relationship between method choice and ability to generate data, it might be possible to suggest a mechanism that might cause some methods to generate more transferable insights than others.

[58] Spearman's Rank Order Correlation is more appropriate here than Pearson's Correlation Coefficient, as the data are discrete, ordinal data rather than being continuous (Pitman, 1937).

[59] The descriptive language that is appropriate to different levels of correlation is contested, varying within and between disciplines considerably (Akoglu, 2018). In this chapter, I attempt to use uncontroversial language in the context of development economics and public health, as these are the disciplines most closely associated with Cases One and Two.

In order to do just this, the 40 Case One observations of a pairing of MSMS score and ability to generate data were investigated to ascertain the correlation between these two variables. Figure 7.4, below, represents these findings visually.

Figure 7.4: Association between the Maryland Scientific Methods Scale and ability to generate data for the evaluations in Case One



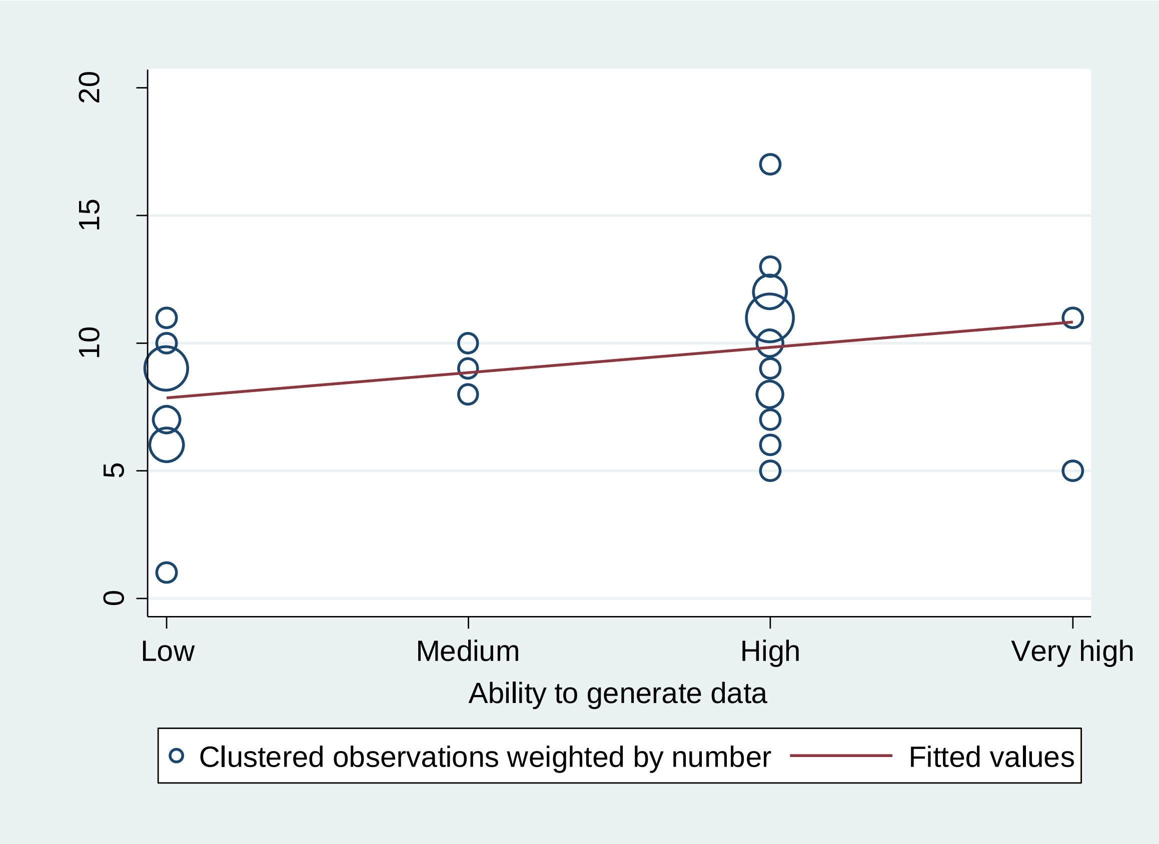
Calculating Spearman's Rank-Order Correlation for these two variables results in a coefficient of 0.2622 with a p-value of 0.1021. As with the previous calculation, the correlation expressed by the coefficient is weak. Also similarly to the previous calculation, the p-value indicates that there is just over a 10% chance that such a result would occur by chance if the hypothesis of no association between the variables were true. Prior theory might suggest that evaluations of RCTs would be considerably more likely to employ expensive custom surveys, whereas quasi-experimental and matching approaches might be more likely to rely on administrative data. However, the data generated though an application of the method described in Chapter Four to show this to only very weakly be the case for evaluations in Case One. Diving into the qualitative data, we can see that this is because many evaluations of the randomised rollout of

Progresa/Opportunidades relied on administrative data rather than surveys that the evaluators themselves devised. This is much less likely to be the case for evaluations of researcher-randomised evaluations, where custom surveys are almost always created in the absence of administrative data that capture outcomes of interest with the required regularity. Total marker scores, then, might be driven by the evaluators' ability to generate data. The next subsection demonstrates this to be the case.

7.1.3 Scores are driven by ability to generate data

In order to investigate the association of total marker score with ability to generate data, the latter variable was encoded as an ordinal variable taking a value from one to four, with one corresponding to 'Low' and four to 'Very High'. Spearman's Rank Correlation Coefficient was then calculated, yielding a coefficient of 0.3758 with a p-value of 0.0219. The coefficient corresponds to a low to moderate level of correlation. The p-value represents a 2.19% chance that such a finding would arise by chance, were there truly no association at all. Figure 7.5, below, represents this information visually.

Figure 7.5: Association between total marker score and ability to generate data for evaluations in Case One

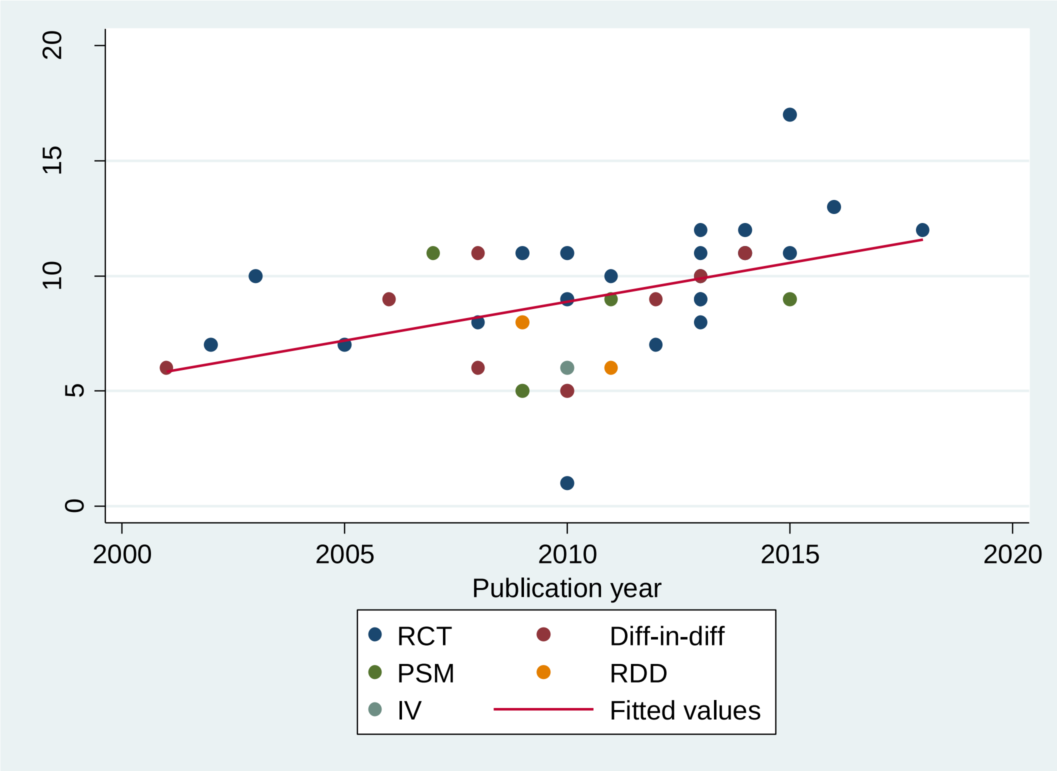


Despite a high level of statistical significance, on its own this correlation would not represent a compelling argument for a causal link between an increased ability to generate data and an increased ability to generate transferable insights. However, it represents powerful *confirmatory* evidence of a causal link for which an obvious mechanism exists. As described in Chapter Five, Section Four, Evaluations that employ custom surveys at baseline and endline score ‘High’ or ‘Very high’ on ability to generate data. Such evaluations should be expected to be better able to generate and report data corresponding to the various contextual features and disaggregations of outcome data that represent 60% (12 of 20) of all the causal markers for evaluations in the set. Evaluations that scored ‘Low’, by contrast, were based on secondary reinterpretation of administrative data that would be less likely to permit the generation and reporting of these data. Evaluations that scored ‘Medium’ employed custom surveys only at baseline. The correlation observed in the data between ability to generate data and total marker score is consistent with the observation that using custom baseline and endline surveys is a necessary condition for the investigation and reporting of all the contextual factors of relevance to an argument for the transferability of findings. To the extent that an evaluation doesn't meet this high standard for ability to generate data, it will be correspondingly limited in its ability to report MICCs and therefore generate transferable findings. This is hardly surprising, but it is another piece of confirmatory evidence that the data generated by the application of the method is generating data that makes sense rather than just noise. It is also interesting and informative in its own right; suggesting that investments in instruments like custom surveys are more important than method choice for facilitating arguments for the transferability of findings.

7.1.4 Scores are improving with time

There is another way in which the data generated by the method described in Chapter Four as applied to Case One appears to be informative. This is the emergent stylised fact that total marker scores are positively correlated with publication year. Figure 7.6, below, represents this association visually.

Figure 7.6: Association between total marker score and publication year for evaluations in Case One



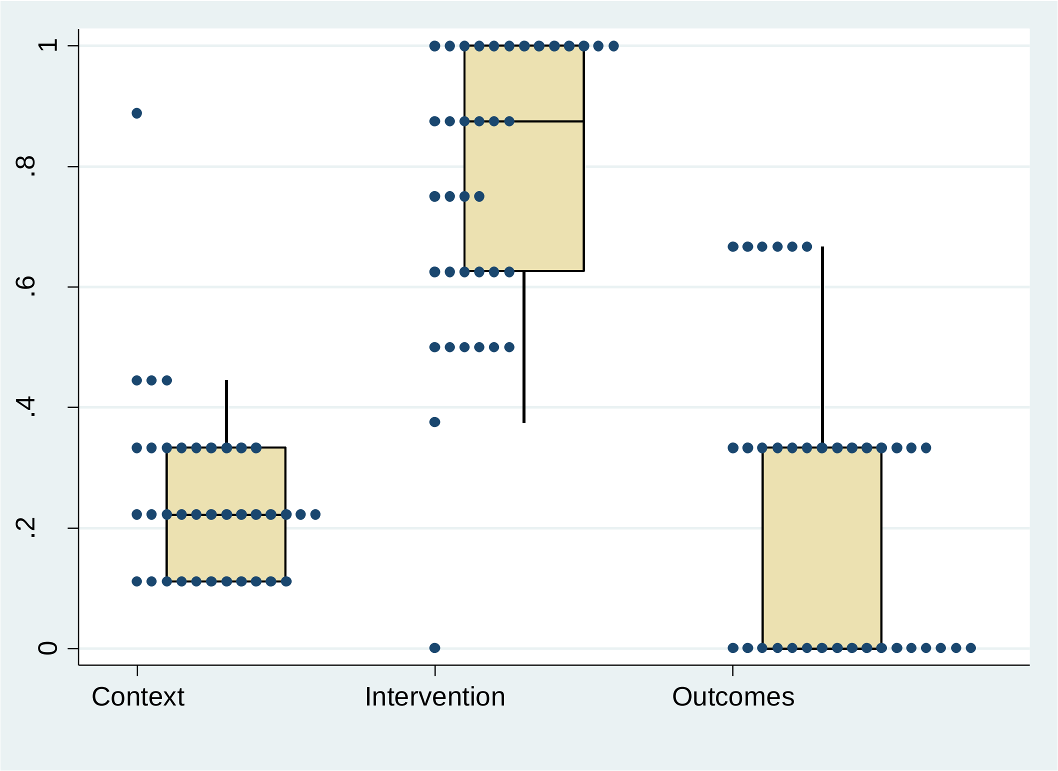
Points on Figure 7.6 are colour-coded by intervention method, to reveal that method choice is not strongly associated with publication year and can therefore be said not to have changed much over time for Case One. Calculating Spearman’s Rank Correlation Coefficient yields a coefficient of 0.5468 with a p-value of 0.0005. This represents a moderately strong correlation that would be extremely unlikely (with a probability of 0.05%) to be observed by chance, were the two variables not in fact associated. This correlation is strong evidence for the success of the method, as it suggests that total marker scores are not generated by some meaningless, random process. If marker scores were merely noise, they would not be expected to exhibit a linear relationship with time. The best explanation of the correlation calculated above is that total marker scores are reflective of some process that is changing with time. This is consistent with the claim made in Chapter Four that marker scores are indeed meaningful, reflective of the ability of evaluations to facilitate arguments that results are transferable to other contexts. The improvement in marker scores with time does not tell us anything about why this is happening. However, in Chapter Eight it will be argued that this correlation is also highly consistent with the emergence of a tendency in the evaluation of development interventions towards more ‘theory-based’ approaches. In that chapter it is argued that practice is improving because evaluators are increasingly seeking to conduct theory-based evaluations, and that this represents a favourable environment for the adoption of realist tools to improve transferability further.

7.1.5 Evaluations report intervention markers much better than outcomes or context markers

The analysis of the data generated for Case One now moves beyond the total marker scores generated for each evaluation, to consider whether it is informative to disaggregate those scores by groups and subgroups of markers. If so, perhaps specific insights can be generated into the areas of focus that would permit evaluators operating within the case to improve the transferability of their results. This subsection generates such insights, providing further evidence of the informative nature of the method, and thereby providing further evidence of its success.

The first level of disaggregation considered is between groups. Figure 7.7, below, shows that the distribution of scores for each the three different groups was very different. Scores for each group have been expressed as a proportion of the maximum possible score for that group in order to facilitate comparability.

Figure 7.7: Case One group scores reported as proportion of maximum possible score using box plot with individual observations plotted



It is clear from Figure 7.7 that intervention markers were much more likely to be reported than context or outcomes markers for evaluations in Case One. The figure shows that the upper quartile of the distribution of intervention scores is identical with 100%, meaning at least one quarter of evaluations reported all intervention markers. The median score is also much higher than for the other two groups. This suggests that intervention markers are easier to report or that evaluators are more motivated to report them. Consulting the list of markers reveals the former certainly to be true. All evaluators have access to all of the information required to report intervention markers. Not reporting this information is either a choice related to its perceived relative unimportance and perhaps to the constraints of article length for those evaluations published as articles, or it is an unconscious omission. Consultation of the qualitative information collected reveals that evaluators are also highly motivated to report intervention markers. This is because several of the evaluations are cast as primarily tests of variation in intervention features. For example, Davis *et al.* (2002) report an evaluation designed to test the relative importance of transfers to women rather than men and of conditionality relative to a lack of conditionality. Similarly, Benhassine *et al.* (2013) report an evaluation designed to test the importance of strong conditionality compared to merely 'labelling' the transfer. Despite the ease of reporting intervention markers, and the fact that many evaluations are highly motivated to report them, one evaluation reported none of them, and one quarter reported less than 70%, suggesting that very low-cost progress could be made for this group of markers for future evaluations.

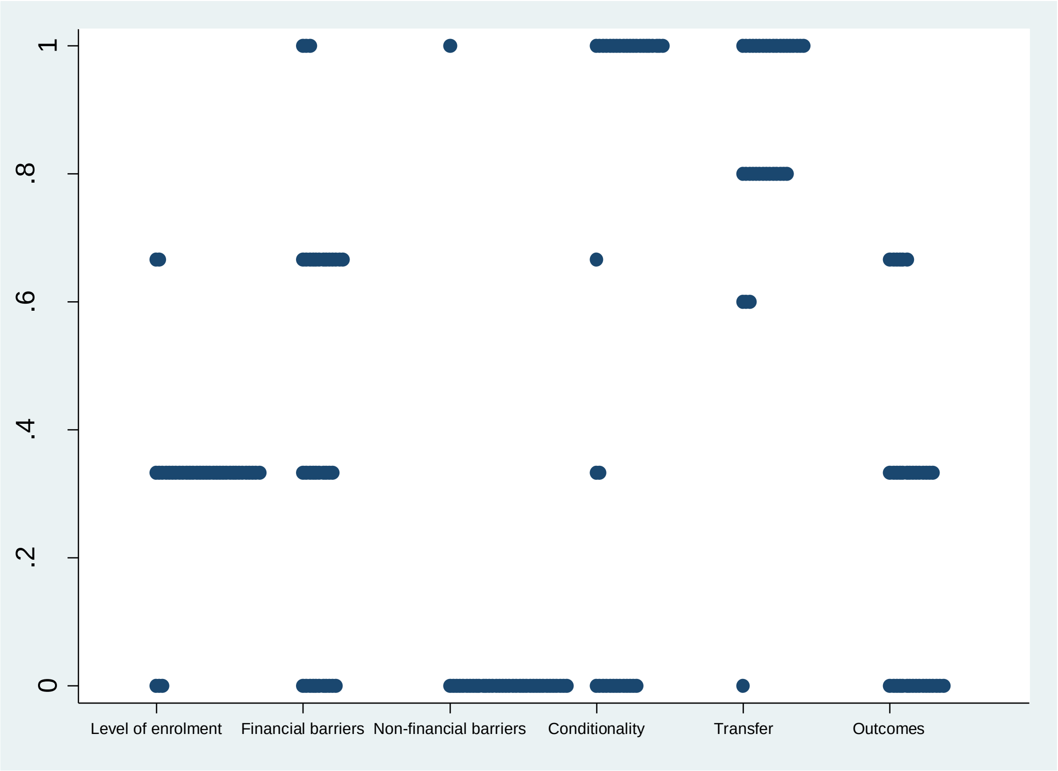
Context markers were unlikely to be reported, with only one evaluation reporting more than 50%, and the median proportion being less than 25%. Despite this, one evaluation (Benhassine et al., 2013) reported eight out of nine, or 89%. This suggests that while most evaluations struggle to report contextual features, or are not motivated to, a huge amount of progress is possible using the tools currently available to researchers.

Outcomes is the only group for which many evaluations reported none of the markers of relevance to the transferability of results. This is further investigated in Subsection 7.1.6, where individual markers are assessed, including outcomes markers. In addition, the results for the context and intervention groups can be better understood by drilling down to the subgroup level. This is done in the following subsection.

7.1.6 Desired levels of enrolment and non-financial barriers to enrolment are badly under-reported

The analysis can now turn to the subgroups of markers that make up the main groups. Figure 7.8 presents the distribution of subgroup scores for Case One, below.

Figure 7.8: Case One subgroup scores reported as proportion of maximum possible score using dot plot



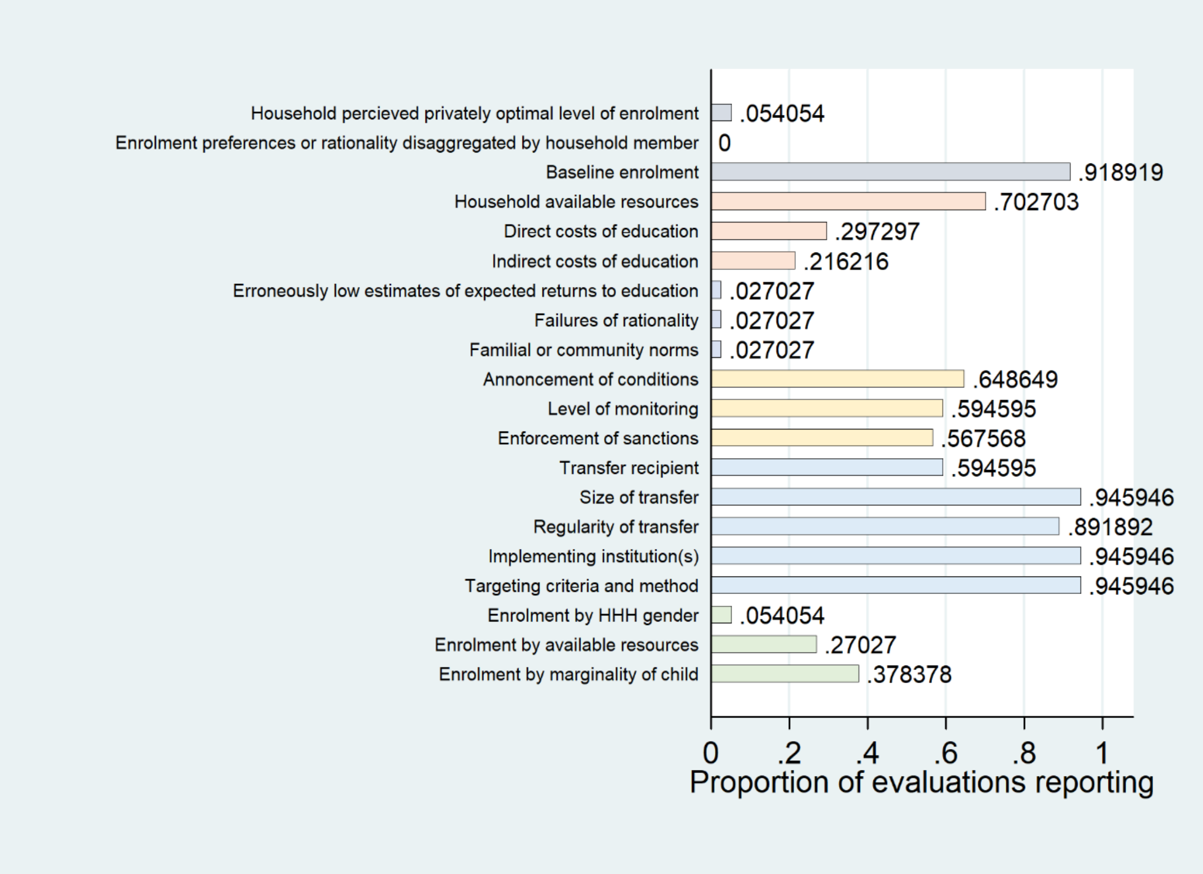
No box plot has been presented for the subgroup scores as it brings more confusion than clarity to the data. This is because, as the dot plot makes clear, the data are much more clustered on a smaller range of values than for the groups. This means, among other things, that the median score is identical with one or more of the first and third quartile scores for five out of the six subgroups, rendering a box plot confusing. The dot plot makes very clear that low context group scores were being driven by the subgroups corresponding to level of enrolment and particularly non-financial barriers rather than the financial barriers group. Disaggregating the intervention group, it is clear that markers in the transfer subgroup are better and more reliably reported than those in the conditionality group, with many zero scores. However, conditionality subgroup scores are more likely to be 100% than those for the transfer subgroup, meaning that interventions that do report markers of conditionality often report them all.

These insights are informative for evaluators working within the case as well as those seeking to understand the literature as a whole or synthesise its findings. For instance, they tell evaluators that although non-financial barriers to enrolment are widely understood to drive the effectiveness of the conditionality mechanism, these barriers are under-studied, and we have almost no evidence base within the specific case of CCTs for enrolment to support that theory. A reanalysis of existing CCT evaluations supported by a survey of non-financial barriers to enrolment in the populations studied could therefore enrich the literature immensely. Likewise, a new evaluation of a CCT program designed to promote school enrolment could be designed to test the relative importance of the conditionality mechanism in the presence of strong and/or weak non-financial barriers, or different distributions of different financial barriers, and thereby contribute immensely to middle-range theory in this area. The previous subsections have argued that the marker score measure is likely to be meaningful rather than being generated by noise. The evidence of this subsection has corroborated this finding and strengthened it by demonstrating some of the specific insights generated. Together, these subsections are strong evidence that the method has been successful. More evidence for this claim can be generated by turning to insights generated at the single marker level and at the level of single evaluations. The following subsections do this.

7.1.7 Many markers are under-reported despite the ability of evaluations to generate the data required

This subsection begins by reporting the proportion of evaluations reporting them for markers of transferability in Case One. Then, individual markers are interrogated to generate insights into the reporting of those markers for evaluations in the set. These insights are informative both about current practice and about the potential for evaluations to generate more transferable results. Figure 7.9, below, reports the proportion of evaluations reporting for each marker in the case, with markers colour-coded and ordered by subgroup and subgroups ordered by group. Some of the same information as was discussed in the previous subsections is visible. For example, markers in the non-financial barriers subgroup are uniformly rarely reported. However, new information is also revealed by drilling down the individual marker level. For the first time, we can see that enrolment preferences or rationality disaggregated by household member have not been reported by any of the evaluations in the set. It is also clear that the household's perceived privately optimal level of enrolment has rarely been reported, and that scores for the Level of Enrolment subgroup are driven by the reporting of baseline level of enrolment, which almost every evaluation reported.

Figure 7.9: Proportion of evaluations reporting for all markers in Case One



Markers are colour-coded by subgroup using the same colours as Table 6.4. They are ordered by group and subgroup.

Another surprising finding is that transfer recipient is only reported by 59% of evaluations despite the central importance of the empowerment mechanism in the literature. This means that in 41% of cases, it is challenging to make an argument for the transferability of results from an evaluation because the recipient is not clear from the evaluation report.

Looking across the subgroups and grouping markers instead by the mechanism to which they are relevant, it is clear that the determinants of the empowerment mechanism are relatively under-studied for this case. In order to investigate the extent to which the empowerment mechanism is driving the effect on school enrolment of a CCT, an evaluation could merely test two versions of the CCT; one in which transfers are made to heads of households, and another where they are made to women. Armand and Carneiro (2018) and Benhassine *et al.* (Benhassine et al., 2013) do just this. However, in order to make an argument for the transferability of the result of this test to any other population, the markers associated with *how* and *why* the empowerment effect was activated would need to be reported. It is heartening, therefore, that Armand and Carneiro (2018) is one of the two evaluations that do report enrolment disaggregated by household head gender and that Benhassine *et al.* (2013) is the only evaluation that reports information about familial or community norms. However, it is unfortunate for the cumulation of knowledge about the relevant middle-range theory that neither of these two evaluations report enrolment preferences or failures of rationality disaggregated by household member.

It might be argued that the very low levels of reporting for some markers suggests that they are too challenging to report. However, all the markers have been reported by at least one evaluation, with the exception of enrolment preferences or rationality disaggregated by household member. This marker is reportable, though admittedly at considerable expense. The level of expense required is that level associated with a survey of recipients with separate units for household members. This is a demanding level of expense but not unheard of. Attanasio *et al.* (2010) employ such a survey, for example. The rest of the context markers can be generated and therefore reported using baseline household surveys of the type employed by most evaluations in the set. The intervention markers can be reported on the basis of process data that should be available to all evaluators. Two of the three outcomes markers require endline data that can be disaggregated by household and matched to household characteristics; one of them requires endline data that can be matched to individual children. This is challenging, but household-disaggregated endline data is available to most of the evaluations in the set, and many have access to endling data disaggregated by child. Therefore, very few of the markers are too challenging to generate and report for the majority of evaluations in the set.

We can give an explanation of low levels of reporting of some markers that better fits the data. Consider the fact that only 16 evaluations in the set report outcome data disaggregated by household resources despite the fact that 26 evaluations report household available resources. For at least 10 evaluations in the set, it would have been simple to report outcome data disaggregated by household available resources. These evaluations did not do so because the evaluators did not consider it important. The inference to the best explanation in this case is that these evaluators were not attempting to distinguish between the operations of different mechanisms in their context. Rather, they were treating the intervention as a package product to be tested for the specific context. They had not thought systematically about what was required to facilitate an argument for the transferability of those results to other contexts. The method being assessed here for specifying the requirements of transferability and testing evaluations against them would have provided those evaluators with a means to improve the transferability of their results at no extra cost. This demonstrates the usefulness of the method in assessing the state of the evaluation evidence. It also demonstrates the potential usefulness of a realist approach for evaluators seeking to generate transferable insights.

It might also be argued that very low levels of reporting for some markers suggests that they have been mis-specified and are not in fact relevant to an argument for the transferability of results to some other context. Chapter Five has described the argument that led to the inclusion of every marker in the set. Some of these markers may be contested by other domain specialists in CCTs for enrolment. We would not expect the model of intervention causation that is summarised in Chapter Six to remain unchanged but rather we should expect it to change and improve in response to further analytic and empirical work. It may therefore be the case that some markers can be replaced by others, perhaps ones that are easier to report. However, Chapter Five argues that the list of markers for Case One is the correct list required by the theory of intervention causation that currently predominates in the literature. An argument to the contrary would require a similarly detailed analytic project to that described in Chapter Five.

It is also surprising that only a slim majority of evaluations report each of the conditionality markers, when these can be generated from process data that is available to all evaluators. This is especially surprising given that it is universally acknowledged that the level of announcement, monitoring and enforcement of conditionality should be an important determinant of the level of action of the substitution mechanism. Despite this, 13 evaluations (35%) reported none of the conditionality markers. By applying the systematic approach to thinking about transferability provided by the method being assessed here, these evaluators would have been able to facilitate an argument for the similarity or difference of the intervention they studied to other interventions either implemented or under consideration. This would have enormously increased the usefulness of their results to other researchers and to policy makers. This is further evidence that the method being assessed has generated informative insights for the Case and is therefore a success. That evaluators are currently failing to report features of intervention context that are easily available to them confirms the excess of successionist thinking among evaluators suggested by the literature review. It also confirms that a more thoroughly theory-based approach to evaluation is sorely needed and that realist programme theory generation provides a framework for that effort.

7.1.8 The method facilitates the systematic critique of individual evaluations in the case

Applying the method described in Chapter Four to Case One has generated informative insights across evaluations. Examining individual evaluations, it is also clear that the method provides a useful analytic lens. Benhassine *et al.* (2013) has already been mentioned several times in this Chapter as an example of good practice. This evaluation reports 17 of the 20 markers, facilitating arguments for the transferability of its findings. However, even for this evaluation, the transferability of findings could have been improved by generating a list of required markers like that described in Chapter Six and reporting them. One of the markers not reported could not have been generated using the surveys developed by Benhassine *et al.* (*ibid*), and altering their survey instruments so as to contain individual household member units might have required resources that were not available to the evaluators. However, the two remaining markers not reported were disaggregated outcome data by gender of household head and by household available resources. These could have been reported, and would have helped to facilitate even stronger arguments for the relative action of different mechanisms in the setting than are possible with the markers reported by the evaluators.

Several evaluations displayed a sensitivity to middle-range theory in the full text, but nevertheless failed to report the markers that would have facilitated an argument for the action of mechanisms in the study context and thereby an argument for the transferability of findings. For example, Chaudhury *et al.* (2013, p.22), write that it ‘could be that the direct costs and opportunity costs of schooling may be considerably higher for older children’ and suggest that this could explain lower levels of enrolment for older children. However, they only report direct costs to education, despite conducting a custom household survey that could have asked families about their perception of the earning potential of children in their community, even if more reliable data were not available. This would have helped to confirm or reject the importance of the substitution mechanism relative to the income mechanism in the study context and thereby provided premises for an argument for the transferability of findings to other contexts.

Baird *et al.* (2010) is an interesting example of an expensive evaluation employing custom surveys at the household and individual level that collected a large amount of often very specific data, and yet reported very few markers. Part I of the survey used collected information from households including ‘dwelling characteristics, household assets and durables, consumption (food and nonfood), household access to safety nets, and shocks (economic, health, and otherwise) experienced by the household.’ Part II was administered to girls included in the evaluation and generated data including ‘her family background, her education and labor market participation, her health, her dating patterns, sexual behavior, marital expectations, knowledge of HIV/AIDS, her social networks, as well as her own consumption of girl-specific goods (such as soaps, mobile phone airtime, clothing, braids, sodas and alcoholic drinks, etc.)’ (*ibid*, p.61). Despite this extensive data collection, the only context markers reported by the evaluation documents are baseline enrolment and household available resources. Some markers, like direct costs of education, may well have been generated but not reported. All markers could have been generated and reported. Despite this, half were not. Therefore, any argument for the transfer of results to other contexts will rest in large part on implicit assumptions about the equivalence of the causal structure of the intervention as implemented and of the study context to those causal structures in some target context. The evaluators invite readers to make this assumptions, but do not defend them, when writing (*ibid*, p.67) that '[t]he evidence presented in this article provides impetus for the expansion of CCT programs (which already cover much of Latin America) to sub-Saharan Africa.'

7.2 WHAT HAS BEEN LEARNED ABOUT EVALUATIONS IN CASE TWO?

This section discusses the scores generated for evaluations in Case Two by the quantitative interpretation of the dataset constructed as described in Chapter Seven. In addition, these insights are triangulated and enriched though the incorporation of the qualitative information generated. The insights described are further evidence of the success of the method described in Chapter Four. The section follows the same structure as the previous section. The analysis begins at the level of total marker scores assessed across the whole case and becomes progressively more fine-grained to finally consider the generation and reporting of particular markers by individual evaluations.

As for the previous section, before beginning the analysis I represent for convenience the list of MICCs organised by group and subgroup. In this case, this means reproducing Table 6.5.

Table 6.5: Deworming MICCs

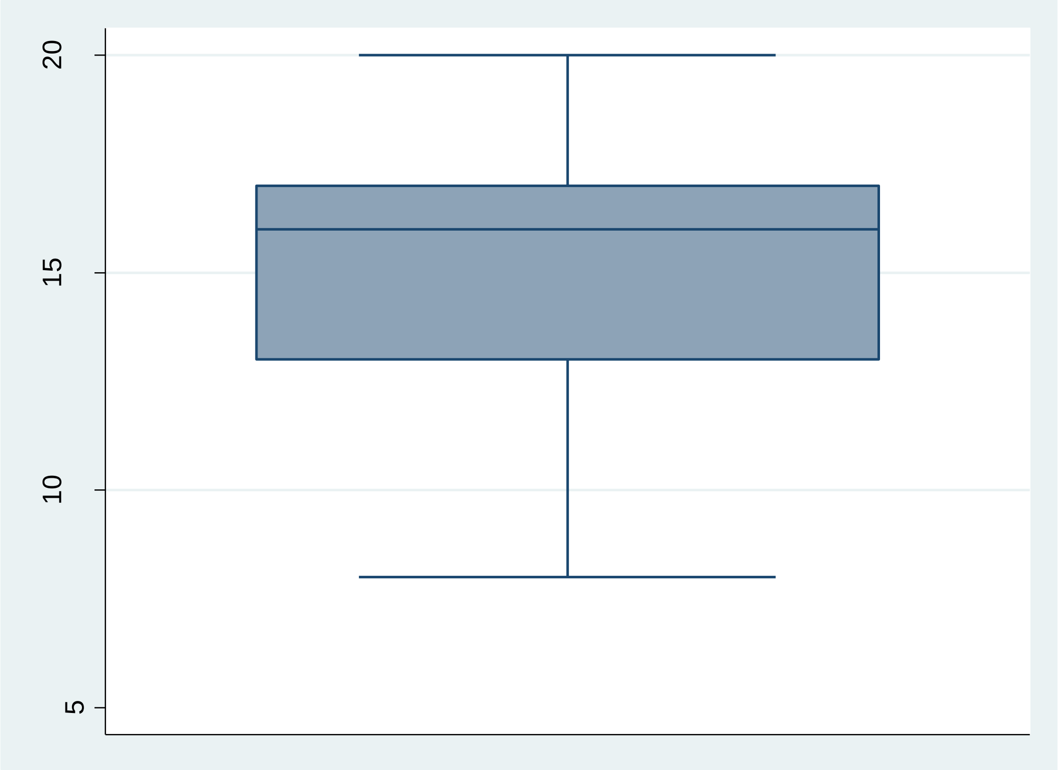
Group	**Subgroup**	**Marker**
Context	Environment context	Hygeine practices
		Sanitation facilities
		Child footwear
		Water sources
	Biological context	Diet quality
		Burden of other disease

	Baseline non-worm	Baseline nutritional status
		Baseline age
	Baseline worm	Baseline infection prevalence
		Baseline infection prevalence by species
		Baseline infection intensity
		Baseline infection intensity by species
Intervention		Drug administered
		Dose
		Frequency
		Targeting (MDA/at risk/diagnosed)
		Implementation agency
Intermediate Outcomes		Proportion treated
		Endline prevalence
		Endline prevalence by species
		Endline intensity
		Endline intensity by species

7.2.1 There is considerable variety in the scores generated for evaluations

As in the first case, total marker scores are widely distributed across evaluations. This suggests that, as for the first case, the method has resulted in the construction of a list of markers that can informatively distinguish between evaluations regarding the level of transferability of their results. Marker scores are neither uniformly high, nor uniformly low. As described in the previous section, either situation would be a worrying sign that the method had failed to be informative for this case. If scores were uniformly high that might be because the insights generated through the systematic approach to transferability being assessed had already been incorporated into the literature by some other means. If scores were uniformly low that would mean that evaluators were never willing and motivated to report those markers. This would suggest that they were miss-specified, either by not in fact being entailed by the theory underpinning evaluations in the set, or by not being reportable in practice. A lack of variation in scores would suggest that the method was not able, for this case, to distinguish between stronger and weaker practice in the facilitation of transferability of results, and was therefore not particularly informative for this case. Therefore, scores being well distributed across evaluations in this case is suggestive of the success of the method in generating informative insights for this case. The fact that this holds across both of the cases against which the method has been tested is strong evidence of its usefulness. Figure 7.10, below, shows the distributions of total marker scores across evaluations in Case Two. A box plot is employed in order to facilitate identification of the median (16) and first and third quartiles (13 and 17). There are no outside values.

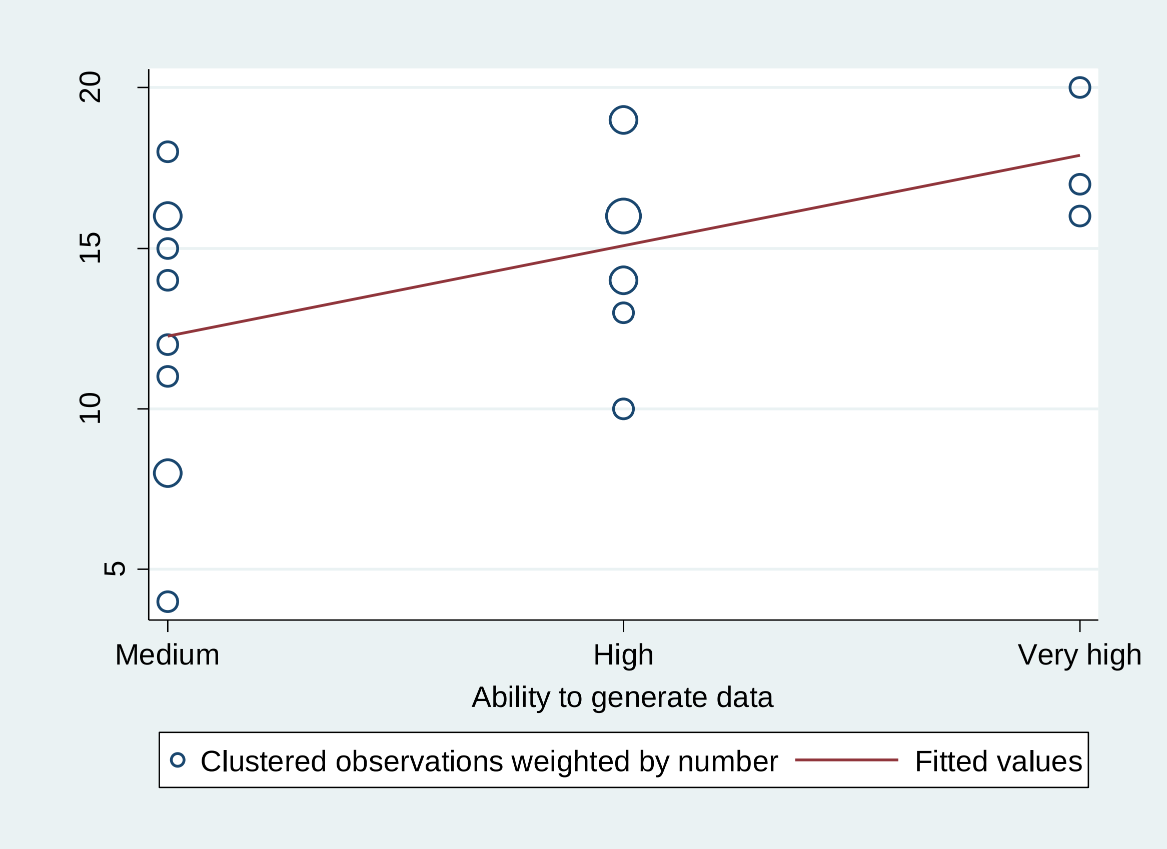
Figure 7.10: Distribution of total marker scores for evaluations in Case Two



7.2.2 Scores are driven by ability to generate data

As all the evaluations in the second case are RCTs, it is not possible to use this case to investigate the relationship between method choice and total marker score. However, it is still possible to investigate the relationship between ability to generate data and total marker score. Encoding ability to generate data as an ordinal variable taking a value from two to four, with two corresponding to ‘Medium’, three to ‘High’ and four to ‘Very high’, it is possible to calculate Spearman’s Rank Correlation Coefficient. This yields a coefficient of 0.5059 and a p-value of 0.0163. This represents a moderately strong correlation that would arise by chance 1.63% of the time, were there in fact no association between the variables. Figure 7.11, below, represents this information visually.

Figure 7.11: Association between total marker score and ability to generate data for evaluations in Case Two

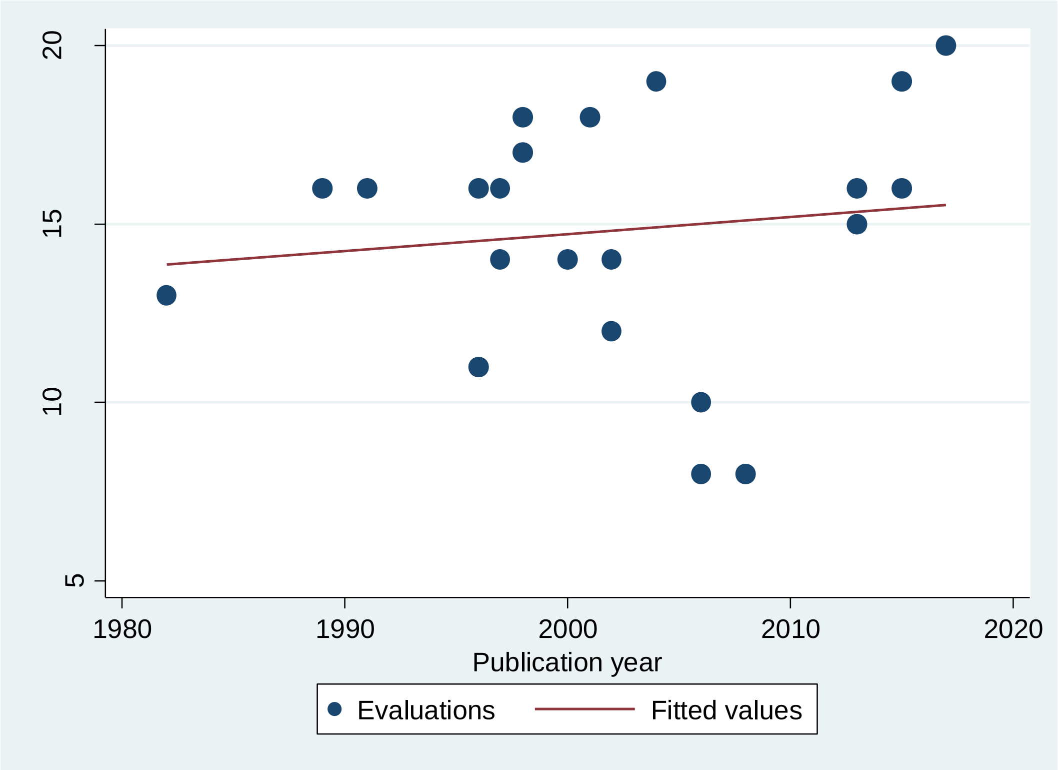


As with the previous case, this correlation is additional evidence that the method as applied has generated a total marker score as well as an 'ability to generate data' variable that are not merely noise, but which represent an underlying kind. This is because mere noise would not be expected to correlate with other noise. This evidence is strengthened by the fact that the strength and direction of this correlation are consistent with existent theory about the properties of evaluations that the 'ability to generate data' variable and total marker score are intended to represent.

7.2.3 Scores do not appear to be improving with time

We might expect that the dataset generated by the application of the method to Case Two would indicate that total marker scores were rising through time, as evaluation practice improves. This is the case for evaluations in Case One. Further, Chapter Eight will argue that this is reflective of a tendency in the (quasi-)experimental impact evaluation of development interventions towards more ‘theory-based’ approaches. However, the data generated in response to Case Two do not support this. Calculating Spearman’s Rank Correlation Coefficient for total marker score and publication year of evaluation for Case Two yields a coefficient of 0.1469 with a p-value of 0.5143. This represents a very weak correlation that is more likely than not (p=51%) to arise by chance if there were no association between publication year and total marker score. This lack of a correlation slightly undermines the claim of the method to have been successful for this case. This is mitigated by the fact that there were relatively few evaluations in the case, with a high variance between total marker scores. This makes finding correlations in the data less likely, even where a true association does exist. It may also be the case that the tendency towards better practice in (quasi-)experimental impact evaluation is stronger for evaluations of CCTs for school enrolment than for evaluations of deworming for child weight. Chapter Eight investigates this hypothesis further. Figure 7.12, below, presents this information visually.

Figure 7.12: Association between total marker score and publication year for evaluations in Case Two



7.2.4 Context scores drive variation in marker reporting between evaluations

The analysis of the data generated for Case Two now moves beyond the total marker scores generated for each evaluation, to consider whether it is informative to disaggregate those scores by groups of markers. As was discussed in the previous section in relation to Case One, investigating this question allows the generation of specific insights into the areas of focus that would permit evaluators operating within the case to improve the transferability of their results. This subsection generates such insights, providing further evidence of the informative nature of the method, and thereby providing further evidence of its success.

Figure 7.13, below, depicts the distribution of marker scores within groups, revealing them to be quite different. Scores for each group have been expressed as a proportion of the maximum possible score for the group in order to render them comparable.

Figure 7.13: Case Two group scores reported as proportion of maximum possible score using box plot with individual observations plotted

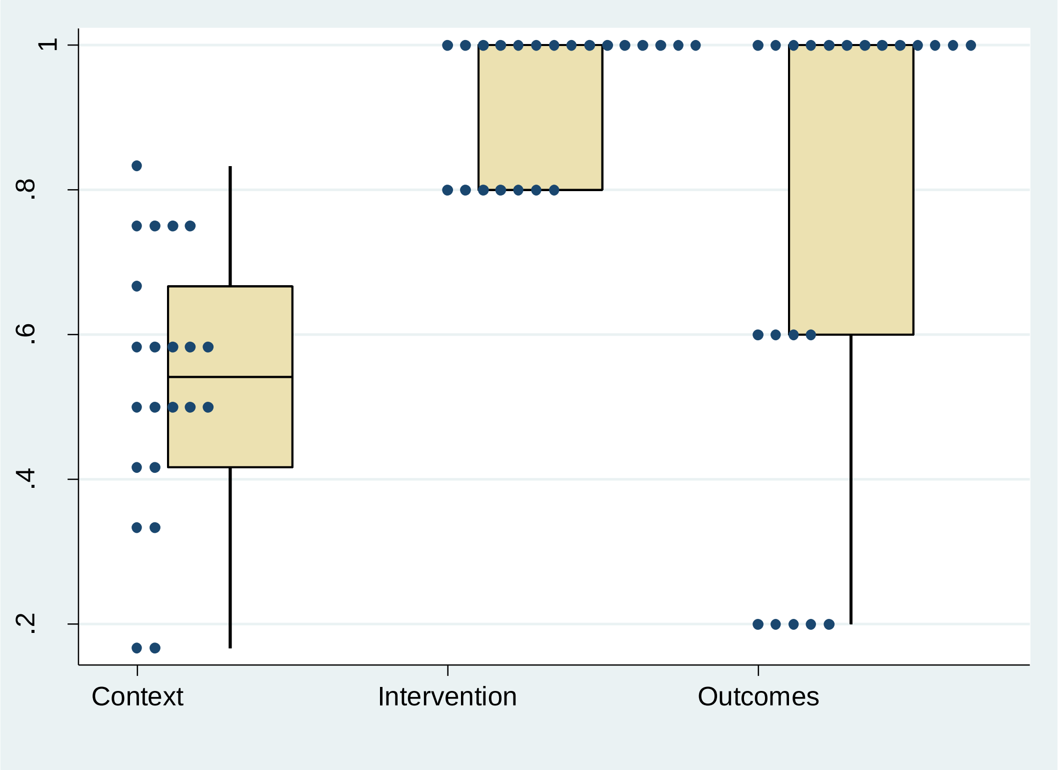


Figure 7.13 makes clear that the variation in scores between evaluations in Case Two is overwhelmingly driven by the group of markers relating to context, with some roll for the

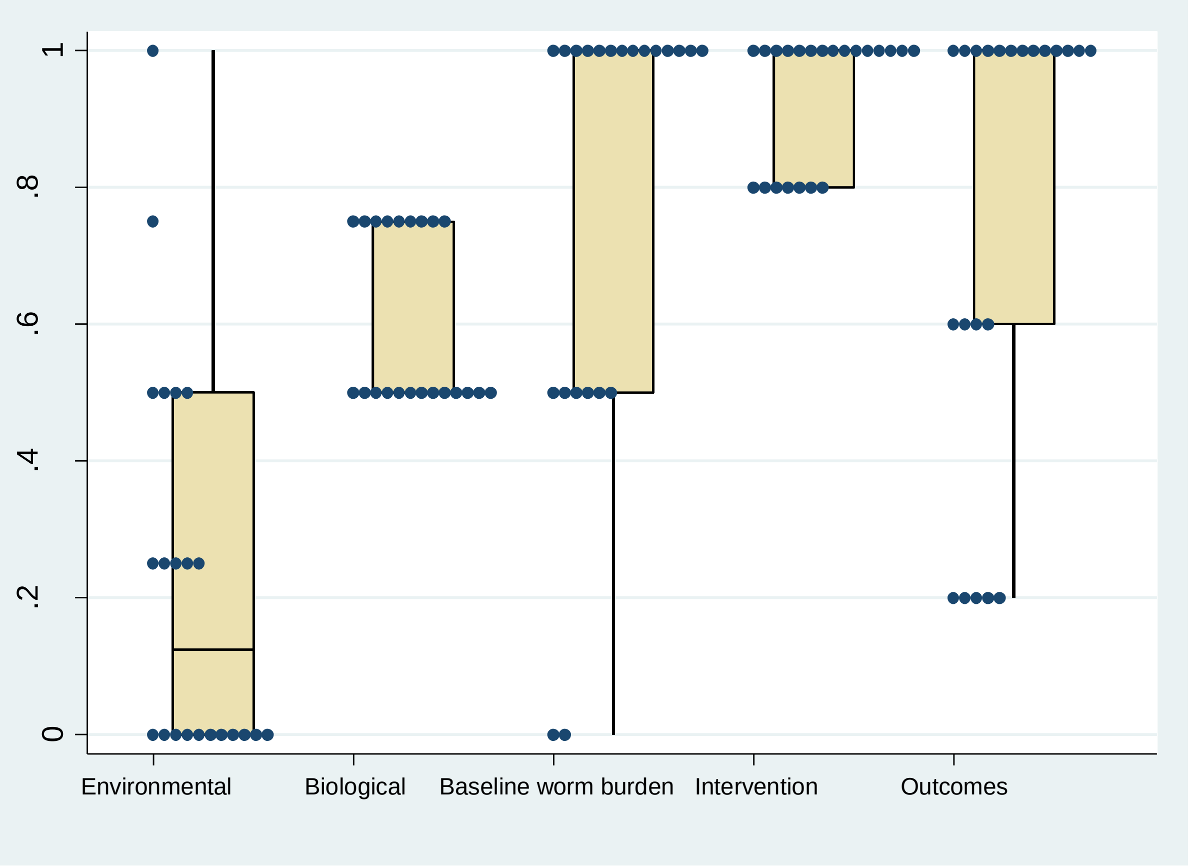
reporting of outcomes. Intervention markers, by contrast are well reported by every evaluation in the set. These results are consistent with Case One. They are also consistent with existing theory in the sense that all of the information required to report intervention markers is available to evaluators in the form of process data.

Outcomes are reported more reliably than for Case One, but it is important to remember when comparing between these cases that the outcome markers for Case One were axes of disaggregation of the primary outcome, whereas this was not the case for Case Two. The theory of intervention causation for Case Two did not require disaggregations of primary outcome in order to facilitate an argument for the relative importance of different mechanisms. Rather, intermediate outcomes were identified that performed this roll. These insights are of some use to evaluators working in the Case and to those seeking to aggregate results across evaluations. They suggest to evaluators that more important and more numerous contributions to theory are possible by focussing on testing the importance of different contextual factors rather than different forms of intervention, as the design of the intervention is clearly better reported across evaluations. This is suggestive evidence that it is better understood. That the method as applied to Case Two has generated these sorts of insights is weak evidence of its success. Stronger evidence is available through a more fine-grained analysis of the reporting of subgroups of markers as well as individual markers. The following subsections present this analysis.

7.2.5 Environmental markers are the least reported, followed by biological and baseline worm burden markers

Analysis of the distribution of scores by subgroup reveals a stark difference between the three subgroups that make up the context group. Figure 7.14, below, shows the distribution of scores by subgroup for evaluations in Case Two.

Figure 7.14: Case Two subgroup scores reported as proportion of maximum possible score using box plot with individual observations plotted

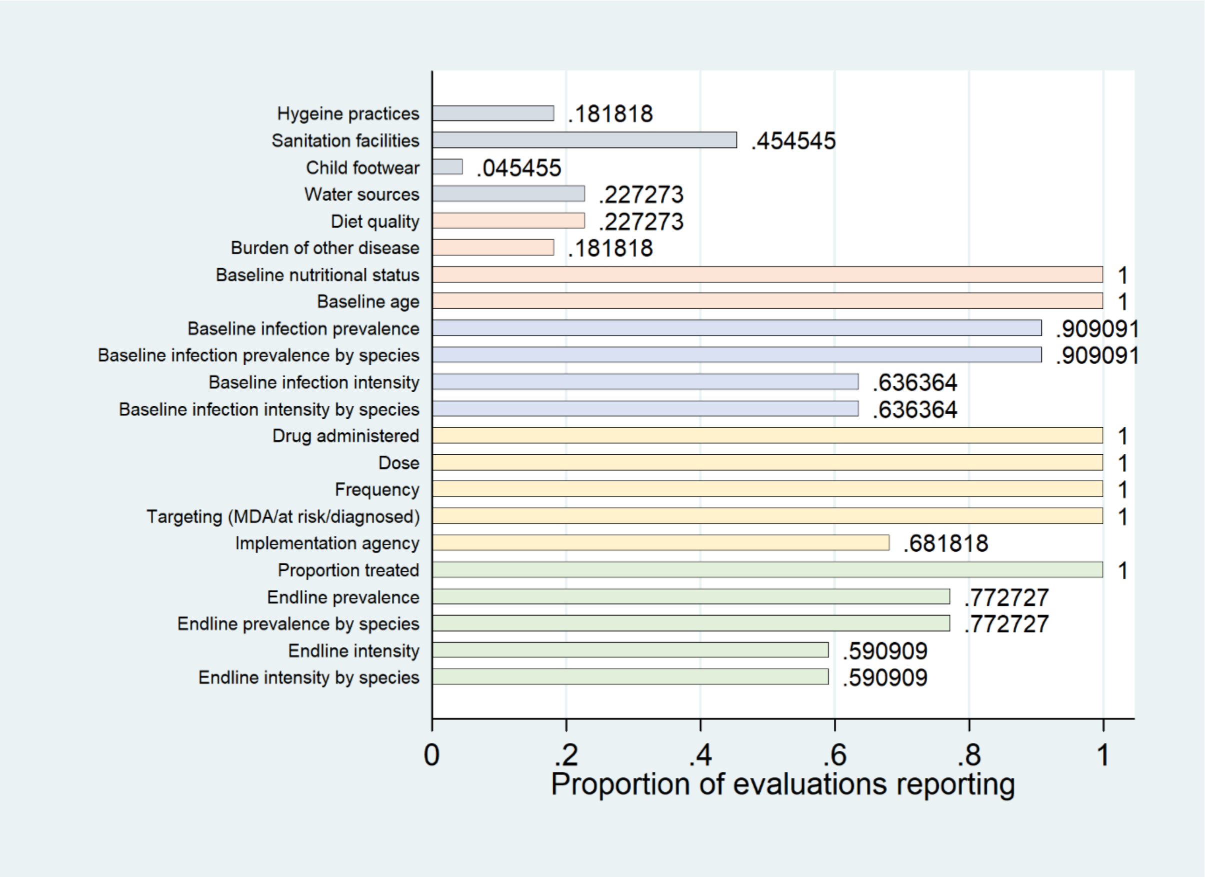


While scores are well distributed for the context group as a whole, half the evaluations in the set report none of the environmental markers, and more than half (14 of 22) report all markers of baseline worm burden. In addition, the distribution within subgroups is informative. For example, despite the fact that 50% of evaluations fail to report any environmental markers, one evaluation in the set reports them all, and six report half or more of them. This finding is informative for evaluators working within the case, as it suggests that the relationship between environmental factors and the effectiveness of deworming for weight is under-studied and also that it is possible to generate all of the key information of relevance to this area of research. This insight will be examined in more detail and rendered more specific in the following subsection.

7.2.6 Many markers are under-reported despite the ability of evaluations to generate the data required

This subsection begins by reporting the proportion of evaluations reporting for markers of transferability in Case Two. Then, individual markers are interrogated to generate insights into the reporting of those markers for evaluations in the set. These insights are informative both about current practice and about the potential for evaluations to generate more transferable results. Figure 7.15, below, reports the proportion of evaluations reporting for each marker in the case, with markers colour-coded and ordered by subgroup and subgroups ordered by group. Some of the same information as was discussed in the previous subsections is visible. For example, markers in the environmental subgroup are on average least likely to be reported. However, new information is also revealed by drilling down the individual marker level. For the first time, we can see that child footwear is the least likely marker to have been reported by evaluations in the set. Sanitation facilities, a marker in the least likely to be reported subgroup, is nevertheless reported by almost half of interventions. We can also see that all variation in the reporting of the intervention group is driven by whether or not the implementation agency or agencies involved were reported by evaluators.

Figure 7.15: Proportion of evaluations reporting for all markers in Case Two



The fact that the child footwear marker is only reported by one evaluation in the set is a key finding illustrative of the power of the method developed in Chapter Four. Entry of soil-transmitted helminths (STHs) into the body via a bare foot is widely acknowledged to be a key vector for STH transmission. Despite this, only one study in the set of evaluations in Case Two reports whether children wear shoes in the community targeted for the intervention. This means that the moderating effect of child footwear may or may not be responsible for much of the effectiveness of deworming interventions. The fact that almost no evaluation reports this key moderating variable for the effectiveness of the intervention suggests that the evidence base could be enriched substantially by revisiting existing evaluations and attempting to generate this data for the communities targeted by the interventions studied. It also means that experimental testing of the moderating effect of footwear on the effectiveness of deworming should similarly enrich the available evidence base. Similarly, some evaluations in the set have paired deworming with an education intervention to attempt to increase its effectiveness, though none have experimentally tested education related to hygiene or footwear, suggesting that such a test could also be high impact.[60] Either research question could lead to a high-impact piece of research useful to those seeking to implement deworming programmes or to evaluate the available evidence on deworming. The ability of the method developed in Chapter Four to generate such insights is strong evidence of its success.

It might be argued that the low levels of reporting for many of the contextual markers suggest that they are too challenging to report. However, the context marker reporting requirements generated in Chapter Seven for evaluations in Case One could be satisfied by small representative surveys of a subset of the study population. Most evaluations in the set (12 of 22) conduct surveys with all recipients of the intervention. The remaining 10 evaluations conduct anthropomorphic measurements with all recipients of the intervention, at which opportunity a short survey might be able to be administered at a small additional cost. In addition, at least one evaluation manages to report all of the context markers despite no unusual methods being used. As with the first case, the suggestion that contextual markers might be too challenging to report does not hold much weight.

Another set of markers are costly to report. In order to report infection intensities at baseline or endline, faecal smear examinations such as the Kato-Katz technique must be employed, which are costly, at around US$2 in the Tanzanian context (Speich et al., 2010). Despite this, 59% of evaluations reported infection intensities at endline and 63% of evaluations reported them at baseline. This suggests that when a moderating factor is widely understood to affect treatment effectiveness, evaluators (and funders of research) are often willing to go to considerable expense to report it, at least in the context of Case One evaluations.

7.2.7 The method facilitates the systematic critique of individual evaluations in the case

Exploring the data generated for Case Two at the level of individual observations in the dataset allows for detailed, systematic critique of the methodological decision made. This subsection

[60] See, for example, Zhang *et al.* (2017)

provides two examples to reinforce this chapter's argument for the success of the method developed in Chapter Four. Firstly, consider Awasthi *et al.* (2008). This evaluation is one of the two evaluations in the set that reports the lowest number of markers. This is partly driven by the inability of the evaluators to conduct faecal smear examinations at baseline or endline. Presumably, this was as a result of cost or capacity constraints in what is a large, cluster-randomised controlled trial across 50 villages in the area of Lucknow, northern India. However, the evaluation only reported two of the biological context markers and none of the environmental context markers. This is despite the fact that Awasthi lives and works in Lucknow, northern India, and several co-authors of the article reporting the evaluation have worked on the area previously. These evaluators know the area and have the means to generate descriptive data relating to the features of context of relevance to the transferability of findings. Awasthi *et al.* are careful not to over-claim the transferability of their findings. They write*, inter alia*:

> *'In such urban slums in the 1990s, five 6-monthly rounds of single dose anthelmintic treatment of malnourished, poor children initially aged 1–5 years results in substantial weight gain. The ICDS system could provide a sustainable, inexpensive approach to the delivery of anthelmintics or micronutrient supplements to such populations. As, however, we do not know the control parasite burden, these results are difficult to generalize.' (ibid, p.e223)*

Clearly, Awasthi *et al.* are sensitive to the importance of reporting markers of the causal structure of an evaluation context in order to facilitate the transferability of results. Further, they had the ability to do this more successfully in the case of this evaluation. This suggests that realist programme theory mapping could have been of use to these authors and could be of use to authors like them, in providing a systematic guideline to the assessment of the markers of transferability that an evaluation should report.

Zhang *et al.* (2017) report 20 of 22 markers of intervention causation, demonstrating their concern for the reporting of the markers of contextual structure of relevance to the transferability of their results. They talk extensively about the contextual features which may justify a deworming intervention or render it bad value for money relative to alternatives (*ibid*, p.44). Despite this, Zhang *et al.* do not report diet quality or burden of other disease in the study population. This is despite the fact that their methods include focus grouping to facilitate a deep engagement with study villages about health practices. Even indicative data about community-level diet norms and general burden of disease would help to facilitate even stronger arguments for the transferability of Zhang *et al.*'s results. The fact that available data were not reported that would have accomplished this suggests that Zhang *et al.* may also have benefitted from a systematic way of mapping programme theory in order to uncover the determinants of transferability of results.

7.3 LIMITATIONS AND POTENTIAL LIMITATIONS OF THE METHOD

The previous sections have provided strong evidence of the success of the method. However, some limitations of this analysis should be noted. This section discusses those limitations and considers the ways in which they place bounds upon the legitimate use of the method. Some potential limitations of the method are also considered and argued against.

One limitation of the method relates to the purpose of the evaluations studied. Each case was constructed by identifying a well-studied pairing of intervention and outcome, and then attempting to identify all of the evaluations that had studied that intervention-outcome pair. Then, for the evaluations in the set as a whole and for individual evaluations, the transferability of the results specific to that intervention-outcome pair was assessed. Some of the evaluations in both sets were attempting to answer research questions that rendered the transferability of the results of the effect of the identified intervention on the identified outcome less important to the success of the evaluation. For this reason, an important limitation of the method is that it is only useful for judging the transferability of results relating to the intervention-outcome pair. This may not be a very significant component of the utility of the evaluation studied, so it should not be seen as a method for identifying successful or 'useful' evaluations. These properties of evaluations are driven by many contributing factors. Nevertheless, for most evaluations their success and utility will be heavily influenced by the transferability of the results relating to the intervention-outcome pair(s) studied. Some examples form the two cases illustrate and clarify the point:

The type of evaluation that most clearly can answer its research question without generating transferable results is a trial designed to answer the question ‘can this intervention ever effect outcomes?’ Such a trial might be designed to disprove an assertion that a certain type of intervention can *never* work or can *only ever* have small effects. There are no evaluations in either of the studied cases that are designed with this intent in mind, but such evaluations are an important example of studies that can be successful and useful without generating transferable results. The closest examples in either set are so-called ‘efficacy trials’ that seek to test whether an intervention implemented under ‘ideal circumstances’ can lead to a worthwhile change in outcomes. These trials occupy a space close to one end of a spectrum from ‘efficacy’ to ‘effectiveness,’ where effectiveness trials, by contrast, are concerned with the ability of an intervention to function well under ‘real world’ conditions (Singal, Higgins and Waljee, 2014). Several of the evaluations in Case Two could be argued to lie on the efficacy end of this spectrum, for example Gupta and Urrutia (1982).

It might be argued that efficacy trials need not produce transferable results. However, this would be mistaken. The purpose of an efficacy trial is to prove that an intervention can work well for a relevant population, albeit under close to ideal implementation conditions. In order to use the results of an efficacy trial to make a decision, for example whether to fund further evaluations of the same intervention-outcome pair, it is important to know that the population studied is not a very unusual population. In order to know this in the context of a social intervention, the markers of transferability of relevance to intervention causation in that population must be reported. It might be less important to report features of intervention implementation like the nature of the implementation agency. However, even implementation markers should be reported in order to maximise the contribution of the evaluation. This is because ideas about what constitutes ‘perfect’ implementation of an intervention may be contested or may change with time.

There are many evaluations in the set that were designed to test complicated interventions made up of more than one intervention component.[61] For these evaluations, the intervention identified in the case was only one aspect of the intervention being evaluated. For these interventions, it is more challenging to report transferable results. This is because a complicated intervention will involve several different intervention components acting on several outcomes. Each intervention-outcome pairing may rely on the action of different mechanisms. This means that the total set of markers that should be reported in order to facilitate an argument for transferability for the intervention will be much larger than for a simpler intervention involving fewer mechanisms. However, this does not mean that such evaluations cannot be critiqued on the basis that they fail to report markers of transferability for any of the total intervention’s constitutive intervention-outcome pairs. Testing and reporting more complicated interventions is harder. To do so in a way that generates transferable results requires the reporting of a longer list of intervention markers. That this is more onerous does not make it less necessary.

Some evaluations in the cases were attempting to solve knowledge puzzles related to different outcomes of the intervention and reported the outcome considered by the case in passing. For example, Rubio-Codina (2010), is studying the effect of CCTs on time-use and reports enrolment as an intermediate outcome. In this case Rubio-Codina’s evaluation can still be assessed with reference to the determinants of transferability of relevance to the effect of CCTs on enrolment, but such an assessment will only be of tangential relevance to those seeking to use Rubio-Codina’s main results. The transferability of her findings on an intermediate outcome is important, but does not guarantee the transferability of her main results, for which further argument will be needed.

[61] Most notably, in Case One, many evaluations are of Progressa/Opportunidades, a complex intervention of which the education-targeting CCT is only one part. Behrman et al. (2004) is one example. There are also evaluations in Case Two that assess deworming in combination with other interventions. Examples include Zhang et al. (2017) who assess deworming with and education intervention and Ebenezer et al. (2013) who assess deworming with iron supplementation.

Another limitation of the method as deployed for Cases One and Two is that it was not possible in either case to extract a rule of combination for markers from the programme theory map. Chapter Four argued that such a rule was necessary for a useful understanding of the causal structures of relevance to intervention causation. In other words, we must know how important contextual features and intervention features are to producing which outcomes for whom in order to argue for the *extent* of the transferability of results between contexts. Unfortunately, such a rule cannot be derived from the theory maps generated for Case One and Case Two. Therefore, all that is possible in the analysis of papers is to assess which markers have been reported. It is not possible to weight the relative importance of markers.[62] This is a modelling simplification, as markers do not contribute equally to arguments for transferability, just as contextual and intervention features are not all equally important in determining the causal processes that lead to outcomes. Nevertheless, this simplification is required, because the programme theory that has been shown to underpin interventions in the set is not sophisticated enough to support a rule of combination. This means that many arguments are not possible on the basis of the data generated by this research project. For example, it is not possible to rank evaluations in terms of their transferability. The total marker score should not be understood in this way. Nevertheless, many useful insights are still possible, as the previous two sections have shown.

7.4 RESEARCH SUBQUESTION ONE CAN BE ANSWERED IN THE AFFIRMATIVE

Sections One and Two of this chapter have demonstrated that the method developed in Chapter Four generates informative insights for Case One and for Case Two. This has been shown at the level of the cases as a whole, considering them as complete literatures to be analysed for gaps, trends and knowledge puzzles. It has also been shown that the list of markers of transferability provides a useful lens for the analysis of individual evaluations. Section Three acknowledged some limitations of the method, as well as considering and rejecting other criticisms. This

[62] Chapter Nine, Subsection 9.3.4 describes how directed acyclic graphs could be used to translate CIMO configurations and other qualitative statements of programme theory into models that admit of quantitative specification and analysis. However, the information in both the studied cases is not rich enough to ground such a quantitative model of intervention causation.

section makes an argument that the evidence provided in the other sections of this chapter is a sufficient basis on which to answer research subquestion one in the affirmative.

Before answering the first research subquestion. Let us consider how it was arrived at. The primary research question, motivated in Chapter Two, follows:

> *Can we give a useful, systematic account of the relative merits of evidence generated using different (quasi-)experimental development impact evaluation methods that goes beyond internal validity to also consider the transfer of results to other contexts? If so, how?*

In Chapter Three, this question was split into two parts to allow separate investigation of the possibility of the account and what it would mean for that account generated to be 'useful'. Each subquestion was then given a theoretically rich interpretation as follows:

1. *What are there systematic differences, if any, between (quasi-)experimental impact evaluation methods regarding the extent to which they report on the barriers and enablers of intervention mechanisms present in the study context and the extent to which they report the degree to which different mechanisms are responsible for changes in outcomes?*

2. *For epistemic communities of development experts, what are the shared notions of validity concerning what counts as a 'high quality' (quasi-)experimental impact evaluation? Further, what are the features of these accounts that are valued by members of the community, and what unresolved puzzles or nascent crises undermine them?*

Chapter Four operationalised these questions to create a research protocol for each. In this process, a suitable existing method was identified to answer Subquestion Two. However, no such method exists to answer Subquestion One. Therefore, a novel method had to be designed. Subsection 1.3 of Chapter Four identified an existing tool that might be adaptable to create a method for assessing the transferability of results generated by any method of (quasi-)experimental impact evaluation. This opened the possibility of creating a novel method to answer Subquestion One. In its operationalised form, the investigation of the first research subquestion then became twofold. On the one hand, the interesting research question became ‘can realist programme theory mapping be adapted to create a tool to assess the transferability of (quasi-)experimental impact evaluation results? If so, what can it tell us about the relative merits of evidence generated using different methods, both as they are currently used and as they might be used better in future?’ This operationalised form of research subquestion one is reproduced below for clarity:

1.

a) Can realist programme theory mapping be adapted to create a tool to assess the transferability of (quasi-)experimental development impact evaluation results?

b) If so, what can it tell us about the systematic differences, if any, between (quasi-)experimental impact evaluation methods regarding the extent to which they report on the barriers and enablers of intervention mechanisms present in the study context and the extent to which they report the degree to which different mechanisms are responsible for changes in outcomes…

i. as they are currently used?

and

ii. as they might be used?

Chapter Four laid the foundation for an answer to research subquestion one by operationalising it, developing a method for assessing the transferability of (quasi-)experimental impact evaluation results based on realist programme theory mapping. Chapters Five and Six describe the deployment of that method for two cases, generating two datasets. This chapter has analysed those datasets, showing that they contain insights of relevance to researchers working in the cases and those attempting to synthesise evidence. The success of the method is demonstrated by those insights. As the method developed in Chapter Four has proved highly capable of generating useful insights for both cases, it can be considered a success. Thus, research subquestion one, part a) can be answered in the affirmative.

8 Opportunities to influence practice in the turn towards theory-based evaluation

The previous chapter has argued that the novel method developed in response to research subquestion one generates informative insights and is therefore a success. This is sufficient to answer research subquestion 1. a) in the affirmative: realist programme theory mapping can be adapted to create a tool to assess the transferability of (quasi-)experimental development impact evaluation results. This chapter develops an answer to research subquestion two, recapped along with the other research questions in the box below.

Research question recap

Research question	**Answered in**
Research subquestions:	
1.	
a) Can realist programme theory mapping be adapted to create a tool to assess the transferability of (quasi-)experimental development impact evaluation results?	Chapter 7
b) If so, what can it tell us about the systematic differences, if any, between (quasi-)experimental impact evaluation methods regarding the extent to which they report on the barriers and enablers of intervention mechanisms present in the study context and the extent to which they report the degree to which different mechanisms are responsible for changes in outcomes…	
i. as they are currently used? and	Chapter 9
ii. as they might be used?	Chapter 9
2. For epistemic communities of development experts, what are the shared notions of validity concerning what counts as a ‘high quality’ (quasi-)experimental impact evaluation? Further, what are the features of these accounts that are valued by members of the community, and what unresolved puzzles or nascent crises undermine them?	Chapter 8
Primary research question:	
Can we give a useful, systematic account of the relative merits of evidence generated using different (quasi-)experimental development impact evaluation methods that goes beyond internal validity to also consider the extent to which methods facilitate the transfer of results to other contexts? If so, how?	Chapter 9

This chapter begins with a description of the implementation of the research protocol for research subquestion two, including a description of the sample of interviewees. In Section Two, two epistemic communities are identified in the data generated in response to research subquestion two. Section Three describes the relative lack of attention to transferability paid by interviewees, a possible barrier to the usefulness of the account generated in the following chapter. Section Four describes three nascent crises in the accounts of (quasi-)experimental impact evaluation quality of the communities identified that present opportunities for the account of the following chapter to be useful. Section Five argues that ‘theory-based evaluation’ is an emerging hegemony in (quasi-)experimental development impact evaluation and that very recent and high profile demands for evaluations to ‘build middle-range theory’ represent an opportunity for the findings of this research project to be framed in a way that is useful for practitioners. This insight will inform the argument of Chapter Nine, in which research subquestion 1. b) is answered, and the answers to all the subquestions are drawn together into an answer to the primary research question.

8.1 RESEARCH PROCESS

Chapter Four, Section 4.2 has described how research subquestion two was operationalised to generate the findings presented in this chapter. Before presenting those findings, this section recaps the protocol developed in Chapter Four, describes how data were managed and organised so as to generate the findings, and characterises in broad terms the interview sample and the literature reviewed.

8.1.1 Protocol recap

As Chapter Four has described, the research protocol adopted in order to answer research subquestion two was constituted of two data generation strategies adopted to satisfy three objectives.

Objectives

A. Identify epistemic communities that claim authority over judgements of the quality of development intervention evaluation evidence.

B. Identify their ‘shared notions of validity’ concerning what counts as a ‘high quality’ (quasi-)experimental impact evaluation.

C. Identify the features of these accounts that are valued by members of the community as well as any unresolved puzzles or nascent crises that put pressure on them.

Data-generation strategies

I. Consultation of the literature, both academic and otherwise published by authors claiming authority over what counts as a high-quality evaluation of a development intervention.

II. Semi-structured interviews with experts on what counts as a high-quality evaluation of a development intervention.

8.1.2 Data management and findings development

The findings presented in this chapter are organised by objective, with Section Two addressing objectives A and B, and Sections Three to Five addressing objective C. However, these findings were not arrived at sequentially. Rather, one account that satisfied all three objectives was developed by working on them in parallel as described in Chapter Four, Subsection 4.2.5. This account was arrived at through a retroductive approach that involved moving iteratively between the data generated through both strategies and an account in progress that attempted to satisfy each objective.

‘Retroductive’ is a better description of the research process than would be ‘grounded theory’ because I began the process of consulting the data with theories about the epistemic communities that I might identify, and I moved between that theory and the data, updating and revising the theory as I went and returning to the data with new ways of looking at it. My initial theories were informed by reading conducted during the literature review and upgrade process for this thesis. I suspected that I might find divisions between disciplines, or at least between natural and social scientists, as well as divisions between RCT enthusiasts and sceptics. Some of these theories were maintained and revised and others discarded as described in the next section.

My interviews were recorded and were structured so as to facilitate coding and interpretation of the transcripts. The question prompts used to guide the semi-structured interviews are reproduced in Appendix E. As Chapter Four, Subsection 4.2.5 describes, I took notes during early interviews, but I found these were not useful as they didn’t capture anything not in the transcripts. For later interviews I decided that it was better to focus on interviewing without taking notes except to aid me in remembering follow-up questions that occurred to me as the interviewee was speaking. Once an interview had been conducted, I generated a rough transcript using speech-to-text automatic transcription software, and then went through the text generated to correct (many) errors and to highlight chunks of text of relevance to each of the three objectives listed above.

I used three text files, one for each of the three objectives to keep track of my active hypotheses related to the objectives and to structure notes and quotes relating to these hypotheses. Chunks of text highlighted in the interview transcripts or in readings from the literature were copied into these text files, and had the name of the participant or source appended to them. Where a chunk of text might have been relevant to two objectives it was copied to both text files. Within the text files I used headings for hypotheses such as 'only one epistemic community' or '*randomistas* vs sceptics' and moved or copied chunks of text from interviews and from the literature below these headings, where I also added notes. In this way, with every new interview or every trip back to the literature, hypotheses gained support or contradiction with the adding of new chunks of text or notes. Successful hypotheses emerged in this way as findings, with other hypotheses being gradually discredited or undermined and discarded through the retroductive movement between sources and text files, empirical and theoretical.

8.1.3 Characteristics of the interview sample

Before presenting findings, it is important to describe some characteristics of the interview sample. The 12 interviewees self-described as experts on the (quasi-)experimental impact evaluation of development interventions and were selected on the basis that others had described them in this way. They were selected for the diversity of their institutional affiliations as well as their locations in the Global North or South. Three are development academics working in the Global North. Two work for third sector quasi-academic organisations in the Global North.[63] Two are consultants for private sector evaluation consultancies with offices in

[63] 'Think tank' might be a suitable description, though it is so vague as to be unhelpful. Both organisations consider their role to be one of coordination of evaluation research, dissemination of evaluation best practice, and facilitation of innovation in evaluations of development programmes.

the North and South. One is an evaluation specialist at a multilateral development bank working in the South and North. One is an evaluation specialist working in the South and North for a northern country's development agency. One is an independent evaluation consultant based in the Global South and working with southern governments. Two are evaluation specialists working for an INGO in the South and North. The diversity of institutional affiliations and North/South location of the sample suggest that the purposive sampling strategy of selecting on difference was a success at least on these observable dimensions. Quotes from participants are attributed to participant numbers. These numbers were randomly allocated to participants and do not relate to their characteristics in any systematic way.

8.1.4 Characteristics of the literature consulted

Just as it is important to describe the sample that was recruited as part of the second research strategy, it is important to sketch the contours of the literature that was consulted for the first research strategy. A broad literature exists that analyses the epistemic communities active in the production and use of evidence in international development. Most of this literature, however, is too broad to be of relevance to an attempt to answer the second research subquestion. For example, Eyben et al.'s book *The Politics of Evidence and Results in International Development* (2015) 'critically examines the context and history of the current demands for results-oriented measurement and for evidence of value for money.' In this very wide-ranging account, experts in (quasi-)experimental impact evaluation are necessarily lumped together. The data that concerned me – the shared values and emerging frustrations or puzzlements of impact evaluators – are not examined in very much detail. The same is true of some of the many criticisms of 'experimentalism,' or '(new) empiricism' in international development, which take aim at (quasi-)experimental impact evaluators conceptualised as one group (Kelly and McGoey, 2018). Other accounts identify a subgroup of these evaluators such as behavioural economists (charging them with 'behaviourism') but similarly do not examine their 'shared notions of validity concerning what counts as a high quality (quasi-)experimental impact evaluation' (Berndt, 2015). Rather, they base the epistemic community in a shared belief about what constitutes a high-quality *intervention*.

The literature that engages with (quasi-)experimental impact evaluators as members of an epistemic community defined by shared beliefs about impact evaluation quality and that investigates those beliefs at a high level of detail is quite small. This is why I have relied so much in this chapter on Ogden's (2016) book which specifically takes this approach. The rest of the literature of relevance is mostly accounts of the rise of the *randomistas*. Some of these works concentrate on the 'rhetorical, affective, methodological and organizational strategies' of the *randomistas* rather than engaging with their shared notions of evidence quality or the puzzles that they consider unsolved (Donovan, 2018; Kvangraven, 2020). However, others do focus on aspects of that community's beliefs that are of relevance for an answer to research subquestion two. For example, Webber and Prouse (2018) not only do this for the *randomistas* but also for their opponents within the World Bank. The examination of the Bank gives useful insights into the points of agreement and disagreement between RCT-enthusiasts and other members of DIME and especially the Independent Evaluation Group.

8.2 EPISTEMIC COMMUNITIES IDENTIFIED

Chapter Three, Section 3.3 defined the epistemic communities of interest to this research project as communities that claim expertise over the (quasi-)experimental impact evaluation of development interventions and are united by 'shared notions of validity' concerning what counts as a high quality (quasi-)experimental impact evaluation. In order to identify epistemic communities within the sample of interviewees, all interviewees were prompted to describe what they considered to be the characteristics of a high quality (quasi-)experimental impact evaluation. Participant one gave a typical response when asked what constituted a high quality (quasi-)experimental impact evaluation:

> *'Assuming an impact evaluation is appropriate, so you want to know the effect of a programme, we tend to separate between large-*N *and small-*N *approaches. So large-*N*, there is a large number of units of assignment, so there is scope to randomly allocate units to treatment. In an ideal world, you would be able to do that so that actually, the difference that you observe between treatment and control is that the only thing that's different is that*

> *one group received the intervention and one group didn't. Oftentimes you can't do that, and then there are various quasi-experimental techniques you can do. So one element of quality is in terms of the identification strategy. Increasingly a mark of quality… so you know there are factorial designs that compare different variations of a treatment, but I think it is really important that what it's being compared to is business as usual or what we think at that point is the best available intervention that's available at that point so you know, it's not an artificial evaluation. Then there's policy-relevance of the question and if it's not policy-relevant you shouldn't be doing the evaluation in the first place I would say. And you know strong stakeholder engagement from the start, and that will also help with the identification strategy because you need full collaboration from implementers.'*

In this account, as in all the others that were given, the focus of the answer is on internal validity. Two other elements of quality that are mentioned are the policy-relevance of the underlying question and the level of stakeholder engagement. These elements of quality were also mentioned by other participants. Participant Two responded by initially emphasising the importance of 'horses for courses' and the need to justify a (quasi-)experimental impact evaluation approach, then talked about the components of internal validity before emphasising policy relevance as one aspect of what they termed a 'questions-driven' approach to evaluation. Participant 12 exemplifies this internal validity + approach when they say:

> *'A high quality impact evaluation from my perspective needs to tick a couple of different boxes, not just the methods box, also other boxes, in particular the usefulness to the policymaker or the client, whoever has an interest in understanding whether a certain policy has certain effects or not.*

The 'methods box' referred to here had previously been defined as various components of evaluation design for an experimental of quasi-experimental approach that contribute to an argument for internal validity. 'Usefulness to the policymaker' seems very close to 'policy relevance'. Generally, answers followed this format, talking about internal validity for several sentences, and then adding one or two sentences about other features of an evaluation which are desirable such as policy-relevance or stakeholder engagement. In one case, participant eight, who works in a northern government's development agency, the need to consider how integrated the evaluation should be into an adaptive programming approach was also mentioned.

The accounts of evidence quality thus elicited were remarkably homogeneous. Disciplinary differences were not obvious even between participants with a more natural sciences/medical background and those with a more social science background. However, there is one divide that emerges from the interviews and consultation of the literature that could be argued to run deep enough to merit modelling that community as two distinct epistemic communities. That divide is over gold standard thinking. An extreme case was participant five, who staked out an explicit hierarchy of evidence:

> *'I come at the concept of the hierarchy of evidence very much from that model in medicine where you have systematic reviews at the top and expert opinion at the bottom and then there are various levels between that. So I would think about something with a randomised control group as higher quality than something that didn't have a control or comparison group.'*

This hierarchical approach was explicitly endorsed by three participants. It is present in the first block quote of this subsection, in which participant one talks about 'ideally' randomly allocating treatment. However, three participants explicitly rejected this hierarchy. As participant eight said, 'I think we've moved on a bit from a very RCT-focussed approach, maybe in the 2010s, '10, '11, '12.' Or as participant twelve said, 'there are many different methods that you can use to do an impact evaluation that provides robust and methodologically rigorous results.' Other participants did not explicitly endorse either position, talking about general standards for all (quasi-)experimental impact evaluation methods like the importance of minimising bias, overcoming attribution problems, assessing whether the study is sufficiently powered. These participants might be adherents to either position.

The same divide emerges in the grey literature and in academic papers. On the one hand, the World Bank's influential Development Impact Evaluation Group (DIME) (2015, p.1) declares that 'Although there are many ways to set up a comparison group, the randomized controlled trial, where units are randomly allocated to either treatment or control (comparison group), ensures an accurate and balanced comparison and represents the gold standard for evaluation.'[64] On the other, the terms of reference for a major 2012 report from DFID 'takes the view that there are risks attached to any "assumed superiority" of a limited set of methods' (Stern et al., 2012, p.1). In the academic literature, that divide is even more stark. For example, Chapter Three of this thesis made extensive use of papers from a special issue of *Social Science and Medicine* that was devoted to discussion between Deaton and Cartwright, and many high profile commentators. Deaton and Cartwright (2018b, p.3) lament 'magical thinking' about RCTs and seek to discredit 'gold standard' thinking. Many commentators agree, though the majority of respondents are attempting to defend 'a special place for RCTs in our epistemological framework' (Oakes, 2018, p.58).

The *randomistas* and the 'RCT sceptics' are often described in terms analogous to two distinct epistemic communities.[65] In his introduction to *Experimental Conversations,* an excellent collection of interviews conducted with the leading lights on either side of this debate, Ogden (2016) suggests that the *randomistas* and the sceptics are divided from each other by differences in their theory of how development happens, which he calls their 'theory of change.' He suggests that the *randomistas* tend to believe 'that small changes can matter a great deal, that technocratic expertise is highly valuable, and that individuals within institutions matter as much as the institutions themselves', whereas the sceptics disagree with at least one of these three

[64] Of course, randomisation only ensures balance *in expectation* and balance must be tested in any particular case, but this point is often forgotten. See Deaton and Cartwright (2018a).

[65] As discussed in a footnote in the introduction to this thesis, I adopt the term '*randomista*' despite the fact that some, such as Webber and Prouse (2018, p.4) consider it to be a 'gendered, derogatory term intended to flippantly dismiss experimental economists and their success, particularly Esther Duflo.' I use the term as it is in very widespread use, including often as a self-description. However, it is important to note the criticism and to be clear that I do not intend any derogatory connotation.

premises (*ibid*, p.xxx).[66] A close reading of the interviews contained in the book supports this description.[67]

It might be argued that the identification of two distinct epistemic communities within the development experts surveyed in this research project requires dividing the rest of this chapter into two different accounts and fashioning two sets of outputs from this thesis to appeal to different audiences. This is not the case because of the reason for which research subquestion two is being asked. As Chapter Three, Section Three explained, research subquestion two is being asked in order to find out how the answer to the primary research question can be rendered useful to practitioners. This is because a strong philosophical argument for a methodological prescription is not sufficient to change research practice; the change in methodology needs to be demonstrated to be useful. It is this focus on working out how to make this thesis useful that means it is not necessary to divide this chapter into two distinct account of two different communities. The ways in which the thesis can be useful to both communities are the same as they derive from shared features of those two communities. These three overlapping areas are discussed in the rest of this chapter.

Firstly, most of the ‘shared notions of validity’ that divide those two communities concern the marketing of (quasi-)experimental impact evaluation, and conceptions of the role that impact evaluations do or should play in influencing policy. A good example of this sort of critique is Lant Pritchett’s ‘policy sausage critique’ (Ogden, 2016, p.xxvii). Pritchett argues that the most prominent proponents of RCTs claim to want above all else to influence policy but do not have a realistic conception of how policy is produced or how (quasi-)experimental impact evaluations might influence this process. As the scope of research subquestion two is restricted to accounts of (quasi-)experimental impact evaluation quality, such critiques were not relevant data for an attempt to answer it. Rather, the aspect of practitioners’ perceptions of (quasi-)experimental

[66] This page reference is correct, referring to page 30 of the introduction; it is not a placeholder.

[67] I make extensive use of the interviews contained in *Experimental Conversations* in this chapter as a way of accessing the less guarded thoughts of many luminaries of the RCT movement and their most prominent critics. This allows me to triangulate the insights arising from my sample of interviews against the much higher profile subjects to whom Ogden had access.

impact evaluation quality that is relevant for this research project is the way in which they deal with transferability. In the next section I talk about the ways in which practitioners' accounts of (quasi-)experimental impact evaluation quality are lacking in their attention to transferability, and see transferability questions as a new frontier in (quasi-)experimental impact evaluation. These insights applied equally to both the *randomistas* and the sceptics interviewed. In the literature the sceptics are more likely to pay attention to questions of transferability, though as Chapter Two, Section 2.4 noted, they have not been able to generate much concrete advice about how to generate more transferable results. In the way their notions of (quasi-)experimental impact evaluation quality treat transferability, then, both communities can be described in one account.

Secondly, RCT sceptics and *randomistas* in the sample of interviewees agreed on the problems with current best practice (quasi-)experimental impact evaluation, what Objective C for this part of the research project calls 'the unresolved puzzles or nascent crises that put pressure on [their accounts of impact evaluation quality].' Section 8.4 of this chapter describes these nascent crises and how they should inform an answer to the primary research question.

Thirdly, the 'theory-based approach' to (quasi-)experimental impact evaluation described in Section 8.5 is embraced by both communities. The emergence of the theory-based approach as a rising hegemonic paradigm creates a favourable environment for the reception of insights generated by this research project in both communities.

8.3 ATTENTION TO TRANSFERABILITY IS LACKING

When asked to describe what makes an impact evaluation high quality, only one third of participants mentioned anything related to the transferability of results. Those participants who did discuss the need for a good (quasi-)experimental impact evaluation to facilitate transferability of results saw that criterion as a new way of thinking. For example, participant nine offered the following.

> *'That's where we are now: internal validity is leading the whole process but over the last couple of years it seems to me that we have a new trend led by people like Howard White, some researchers at the world bank as well … are also trying now to pinpoint the need to take context into account.'*

This lack of attention to transferability when discussing what makes a high quality (quasi-)experimental impact evaluation is not as present in the evaluation literature. As Chapter Two, Section Four notes, calls for evidence of 'what works' have generally softened to often include the realist inflection 'for whom, in what circumstances' (Pawson and Tilley, 1997, p.220). The World Bank's DIME group (DIME, 2015, p.1), though it endorses RCTs as the gold standard, insists that (quasi-)experimental impact evaluation should be used not just to test 'what works', but also 'what are the mechanisms that drive impact (that is, how/why does it work)?' As Chapter Three has argued, this is precisely the information that is required to make an argument for transferability.

Some participants saw questions of transferability as being the preserve not of (quasi-)experimental impact evaluation methods, but of other methods such as process evaluation specifically or qualitative investigation more broadly. For example participant five, who endorsed a strict hierarchy of evidence, said, 'I think that qualitative research can help you to unpack mechanisms of action and I think that's quite important for thinking about generalisability.' Chapter Nine, Subsection 2.2 discusses this idea further.

A relative lack of attention to questions of transferability in accounts of evidence quality might be seen as a barrier to the possible impact of an account arising out of this research project, decreasing its chances of being useful to experts and therefore decreasing its chance of being adopted. However, data generated through interviews and from consultation with the literature suggest that problems of transferability are perceived as a nascent crisis in (quasi-)experimental impact evaluation, as the next section argues. If this is the case, then the insights of this research project are perfectly suited for framing as a response to that crisis, increasing their chances of being useful.

8.4 UNRESOLVED PROBLEMS AND NASCENT CRISES

This section discusses the problems that participants identified with the shared accounts of (quasi-)experimental impact evaluation quality that defined their epistemic communities. Transferability was seen as a key problem that needs solving. Prevailing evidence synthesis methods were not seen as adequate to the challenge. Further, the incentive structures within which (quasi-)experimental impact evaluation practitioners find themselves were not seen as conducive to encouraging the facilitation of transferability. The subsections of this section present each of these findings in turn, triangulating them with reference to the literature.

8.4.1 Transferability is unresolved

Many participants described 'external validity' or the 'translation of results' or 'generalisability' as 'a problem.' For example, participant eight described external validity as 'a problem' though they went on to say, 'but I think it's changing, it's been recognised [as a problem], there are different people working on it.' Participant seven explained that 'every time I go somewhere and present a [systematic] review… they will say what about my context? What did the studies from Tanzania say?' This participant also expressed frustration that 'one of the questions I get is "Imagine that I'm the education minister for some country, what would you recommend that I do?" going on to say 'So I guess first of all I would say that it's not my job [as a systematic reviewer] to make recommendations, but I do think that we could do more to facilitate the translation. … We have failed to communicate what evidence-informed decisions making is all about and we have failed to take that to where it should be in our work.' These kinds of concern are widespread in the grey and academic literature, too. Development economist Chris Blattman, in a highly popular post on his influential blog (2016), laments that in the evaluations he helped to run in Liberia 'we did not do one of the most important things at all: set up the research to see if this very local insight told us something about the world more broadly'.[68] Most practitioners would not go so far as Lant Pritchett (Ogden, 2016, p.137) and claim that 'people are no better off in terms of knowledge than they were before [the growth of

[68] This post has 243 replies.

(quasi-)experimental impact evaluation in development].’ However, even enthusiastic RCT practitioners like Morduch are concerned that (quasi-)experimental impact evaluations are often presented ‘as fairly universal findings’ when ‘that’s a problem because in fact all results are specific and conditional to context’ (*ibid*, p.59). The need to address the ‘problem of external validity’ was a point of agreement between Deaton and Cartwright (2018a; b) and most of their supporters and critics in the special issue of *Social Science and Medicine* devoted to the strengths and weaknesses of (quasi-)experimental impact evaluation techniques. As Chapter Two has argued, the identification of this problem in the literature review stage of this research was a key motivation for the primary research question. Analysis of the interviews corroborates this finding of the literature review process.

8.4.2 Inadequate evidence synthesis methods

In the face of widespread concern about external validity, participants did not see existing (quasi-)experimental impact evaluation evidence synthesis methods as adequate to the task of taking account of the context-specific nature of treatment effects. For some participants, this was a matter of synthesis methods such as systematic review and meta-analysis being misunderstood. As participant seven put it, ‘we try to do the reviews in a way that takes into account context and not claim that just because it worked on average, that that is what you should pay attention to. More often than not it’s the heterogeneity that’s interesting.’ For other participants, systematic review and meta-analysis were seen as deficient, at least in the way that they are generally practiced. The opinion of participant three was that ‘Based on my interaction with the evidence synthesis world, and authors there, I think they might not necessarily account very well for I won’t say external validity, but I would say context.’ Participant four suggested a reason for the inadequacy of synthesis methods that was also suggested by two other participants: ‘Development interventions are characterised by small effects that are unstable across contexts, so techniques imported from medicine are not sufficient.’

This last comment leaves open the possibility that, as participant one argued, systematic reviews and meta-analyses will be able to do better in future. Two ways in which this might be achieved are suggested. Firstly, the studies available for early systematic reviews were limited. By building on a richer evidence base, later systematic reviews and meta-analyses are better able to investigate the moderation of intervention effects by context. In the literature, this is a widespread hope (Banerjee and Duflo, 2008; Ogden, 2016). It is also argued that better methods, including mixed-method reviews will improve the ability of reviews to understand context, some good examples of which have already been conducted (Oya, Schaefer and Skalidou, 2018; Waddington, Masset and Jimenez, 2018). However, in order for these strategies to be successful, evaluations need to report the markers of intervention causation in context. In the absence of this marker reporting, qualitative work cannot be complemented by quantitative comparisons, increasing risk of bias. Chapter Seven shows that for the two cases examined, these markers were very rarely reported well.

The interviews with participants demonstrate that they share the perception that reporting in evaluations is often inadequate. Four participants complained that poor quality reporting constrained their ability to use the findings of (quasi-)experimental impact evaluations to try to make arguments for transferability. This may represent a constraint on the ability of better review techniques to provide evidence for the transferability of results. As participant seven, who conducts systematic reviews put it:

> *As a systematic reviewer we go through a lot of impact evaluations where, there are studies where we can't use the outputs because they are so poorly reported. It can be everything from not describing the context, so that will include the context of the problem, to describing in detail what is the intervention. This was the design, what actually happened, what did your process evaluation suggest happened, and you know… just basic things like sample size, means and standard deviations.*

Participant five agreed, saying:

> *'Sometimes it's also not always easy to understand the characteristics of the program. What proportion of the village received the transfer, how it was targeted, what percentage of people ... those kinds of basic description of the intervention is sometimes lacking.'*

The two connected problems described in this subsection, of inadequate meta-analysis and poor quality reporting, may represent an opportunity for the systematic account developed in answer to the primary research question to be useful. This theme is returned to in the following chapter.

8.4.3 Current incentives in impact evaluation are seen as a problem

Another family of problems identified by participants was that of incentives. Participant one identified incentives for data sharing as being insufficient. As they said, 'people want to get as many publications out of the data as possible before they share it, because the academic system doesn't necessarily reward sharing of data in the way that they reward publications.' Participant eleven also identified the strong incentive to publish in particular journals as a problematic incentive. The negative effect that they had noticed was on research priorities, saying, '[evaluators need to take] a more systematic approach to what the evidence gaps are and doing… pointing research in the direction of what is most useful rather than what will enable teams to get a publication in a top five journal.' Participant two identified the process through which evaluations are commissioned as a problem:

> *'When it comes to external validity, thinking about how transferable findings are, ... in reality it's a challenge with the incentives structure currently in place. If I were the evaluation adviser in [SSA country redacted] I would mainly be concerned about getting the right piece of work for my needs in [SSA country redacted].'*

Changing the incentives in (quasi-)experimental impact evaluation commissioning and publishing is outside the scope of this or any thesis, so this crisis cannot inform the account of the next chapter.

8.5 THE THEORY-BASED APPROACH AS AN EMERGING HEGEMONIC PARADIGM

Theory-based evaluation is not a new idea. Weiss (1997, p.41) traces its origins to the 1970s in the professional evaluation community in the Global North drawing on examples from public health, sociology and management. More recently, Howard White has championed theory-based evaluation of development interventions, including by ensuring its permeation into the organising principles of 3ie, the influential coordinating body for evaluation of development interventions of which he was the executive director of for six years beginning with its foundation (Savedoff, 2014).

8.5.1 Most participants have adopted the language of a ‘theory-based’ approach

Six of the 12 participants claimed to use a ‘theory-based approach’ (One, Three, Four, Seven, Nine, Ten). Of the remaining six, two said their approach was always based on ‘a theory of change’ (Five, Eleven). Even those who did not use either of these terms talked about the need to ‘unpack mechanisms of action’ or some similar phrase. Although Participant Four claimed that ‘economists have been very reluctant to adopt [a theory-based approach]’ and that ‘a lot of people still think’ that interventions should be accredited as effective in one context or a small number of contexts and then implemented everywhere, none of the participants interviewed spoke in those terms. Of course, adopting the language is not the same as adopting the practice. Participants who endorse a theory-based approach *in theory* might nonetheless produce work that takes a successionist, black-box approach to causation. Nevertheless, the openness of participants to the language of the theory-based approach is useful information to inform the framing of the argument of the final chapter of this thesis.

This endorsement of the language of theory-based evaluation is also widespread in the literature. Rogers (2007) catalogues its permeation through the professional evaluation community, while White (2009) notices a similar rise in the evaluation of development interventions. What is less clear is how much that endorsement and use of the language reflects an adequate change in practice. Chapter Three, Section Two has argued that reporting the markers of intervention causation that are implied by one’s theory is the minimum condition for adequate facilitation of the transferability of the results of an evaluation. Chapter Seven, Subsection 7.1.4 demonstrates that marker scores for the first case can be shown with great confidence to be increasing with time. This demonstrates that the language of theory-based evaluation is not just lip-service in the Conditional Cash Transfers literature, with more contextual information being reported over time. It is also the case that many of the evaluations in the set seek to explore the working and relative importance of the mechanisms behind the action of conditional cash transfers. For example, Davis *et al.* (2002) report an evaluation designed to test the relative importance of transfers to women rather than men and of conditionality relative to a lack of conditionality. Similarly, Benhassine *et al.* (2013) report an evaluation designed to test the importance of strong conditionality compared to merely 'labelling' the transfer. However, marker scores (and therefore the transferability of evaluations) cannot be shown with confidence to be increasing with time in the second case, as Chapter Seven, Subsection 7.2.3 describes. There are many factors that might explain why the strong correlation between marker scores and time observed for evaluations of CCTs is not observed for evaluations of deworming interventions. Firstly, there are fewer evaluations with a higher variance in marker scores in the second case. Secondly, marker scores as a proportion of total scores began from a much lower base for the first case, with a lower median marker score over all evaluations in the case – 9/20 compared to 17/22 in the second case. Both of these features of the data mean that for the second case, the weak positive correlation between MICC scores and time that is observed cannot be confidently said not to have arisen by chance. Specifically, the observed correlation is 0.1469 with a p-value of 0.5143. This represents a very weak correlation that is more likely than not (p=51%) to have arisen by chance. However, it is worth remembering that absence of evidence is not evidence of absence (Altman and Bland, 1995). The increase in the reporting of MICCs in the first case, as well as the growing number of evaluations designed to test theory are strong evidence that theory-based evaluation is not just a buzzword in the evaluation of development interventions. This evidence is not contradicted by anything in the second case, though it is not a pattern that is strongly evident there.

The approach to transferability argued for in Chapter Three was not described by participants when they were asked how transferability of results might be achieved by (quasi-)experimental impact evaluations.[69] However, for some participants it could be elicited using Socratic questioning. Where that attempt failed, participants would endorse the approach to the minimum reporting required to facilitate transferability, if that approach were put to them. A long quotation from the interview with Participant One follows, which illustrates how this played out in that interview:

> *MJ: We've talked about reporting how the intervention was done in such a detailed way that somebody was able to repeat it. So that's one part of context. ... Is there another part of reporting the context that's important? And if so, what is sufficient reporting of that?*
>
> *P1: Say in education, we want to know what are the kind of, main drivers of poor education outcomes for example in a particular context. And know something about the education system, and culture, and the socio-economic and other characteristics of the children or the teachers. It's fairly basic stuff but often it's not reported.*
>
> *MJ: If I'm a part of the team that's conducting the trial, how should I think about which contextual factors to include and report, and which ones to exclude?*
>
> *P1: Hopefully before you run your trial you have done like a proper problem analysis, so that you know what are the elements of context that are relevant. So I guess part of the issue around quality I think and to ensuring better quality studies in future is to ensure there is more formative work that actually informs the impact evaluation in the first place.*
>
> *…*
>
> *MJ: And then just returning to reporting of context, is it that theory of change that determines which contextual factors we need to report?*
>
> *P1: Partially, but I guess you would also want to know… I think that's context about the population, but then you also would want some context about the system.*

[69] See the interview guide used in Appendix E for an idea of the sorts of prompts that were used in semi-structured questioning.

> *MJ: So, those interactions between the context of the population and the context of the institutions, that should be part of our theory of change for the intervention, right? We should think our intervention will work for reasons which include features of the institutions as well as features of the population, right?*
>
> *P1: Yeah, I guess so. Yeah.*
>
> *MJ: So, do you think it's prima-facie reasonable that that should guide our reporting of which contextual factors matter?*
>
> *P1: Yeah, I guess so. Yeah and I guess that is the idea behind advocating for a theory-informed approach is that you use that kind of as the framework for guiding your data collection, your analysis and reporting. Yeah.*

This interview is fairly typical in that the participant has adopted a theory-based approach in several areas of evaluation design and interpretation, but has not done systematic thinking about the reporting requirements implied by a theory-based approach to transferability. This meant that the approach to the reporting required to facilitate transferability was novel for the participant, but was consistent with their existing thinking about (quasi-)experimental impact evaluation quality. Other participants' reaction to this area of discussion were similar. Participant Eight responded to initial queries about what reporting might be required for transferability with 'I don't know at this point, I would need to read up basically' but responded to a description of using theory to guide reporting of markers with 'Yeah, that makes a lot of sense, thank you.' Responses like these are very encouraging for the prospects of the account that emerges from this research to influence practice. Widespread movement towards theory-based evaluation in the development evaluation community means that experts' existing frameworks are compatible with the insights generated by this research project.

8.5.2 The theory-based approach and realism

'Theory-based approach' is a vague term. The main authors who have attempted to explain and promote the approach have kept their explanations at a pragmatic level, rather than explicitly specifying the ontological and epistemic assumptions that underpin the approach. Weiss' foundational book chapter (1995) and paper (1997) are argued by way of concrete examples rather than philosophical analysis. White's writing on the topic follows this same approach (Carvalho and White, 2004; White, 2009, 2010). White's 2010 paper's abstract confusingly says 'I then consider accusations of being 'positivist' and 'linear', which are, respectively, correct and unclear.' However, the text of the article does not contain the word 'positivist' or 'positivism'. The section dealing with 'linearity' does not endorse a positivist position. In the end, it seems, when editing the paper White decided to continue his policy not to engage with epistemological terminology. Despite proponents of theory-based evaluation not explicitly endorsing realism, the ontology of the theory-based approach is implicitly realist. The account is founded on the need to develop and use 'causal theory' and therefore on a generative account of causation. Much use is made of realist terminology, notably 'mechanism', and papers on the topic cite Pawson and Tilley's (1997) realistic evaluation as one *method* that is an example of a theory-based *approach* (e.g. Stame, 2004). More theoretical papers cite Sayer (1992) and other explicitly realist works as authorities from whose works the theory-based approach derives legitimacy (e.g. Davidson, 2000, p.18).

The participants interviewed exemplified the same approach. Participants who endorsed a 'theory-based' approach described it by way of concrete examples. For example, Participant Three described their approach as follows:

> *'What I use now, in my work, is theory of change. I often use a very detailed diagram, which tries to map out, not only from any existing theoretical frameworks, how the change I expect will occur, but also taking into account, very short interviews with key informants. Once I have that, my causal chain, with all the different paths around it, and also accounting for assumptions … I try to document all this in my theory of change. And once I have that, moving from step A to B, I develop different questions around that, from B to C and so on. It helps me later on to run my data analysis.'*

Most participants employed terms like 'causal chain' or 'mechanism' that implied a realist ontology, though no participants mentioned realism by name. In order to maximise the chances of the account developed in this research project to be useful, it may be best to avoid the terminology of Realist Evaluation. Unfortunately, as described in Chapter Three, Section 3.2.3, later Realist evaluators, following Pawson, have rejected the utility of experiments and too often described Realist Evaluation as an alternative to (quasi-)experimental impact evaluation rather than a complement to it. This antagonistic position makes it much more difficult for practitioners of (quasi-)experimental impact evaluation to embrace lessons drawn from capital Realist Evaluation. Chapter Nine makes an adjustment to the account so far developed to mitigate this worry, focussing on the value of producing and using MICCs, and demonstrating that developing CIMO configurations, as I have, is not the only way to get to MICCs. That liberates the account developed from a necessary association with any unwelcome baggage that might accompany the use of CIMOs and the other machinery of Realist Evaluation.

8.5.3 Middle-range theory

One way in which the value of deriving MICCs from program theory might more readily be accepted by pragmatic theory-based evaluators is as a manifestation of 'middle-range' or 'mid-level' theory. Though none of the participants spoke in terms of middle-range theory, this way of speaking has a long pedigree in the theory of evidence-based policy (Nolan and Grant, 1992). Older literature described 'middle-range' theory by its level of abstraction. Merton (1968, p.39) describes middle-range theory thusly:

> *It is intermediate to general theories of social systems which are too remote from particular classes of social behavior, organization and change to account for what is observed and to those detailed orderly descriptions of particulars that are not generalized at all. Middle-range theory involves abstractions, of course, but they are close enough to observed data to be incorporated in propositions that permit empirical testing.*

More recent writing in sociology describes middle-range theory in opposition to the 'many theories, [which] in their efforts to explain everything, succeed only in explaining nothing.'

These theories are described as being ‘too complex, global, abstract and esoteric’ (Nolan and Grant, 1992, p.218). By contrast, middle-range theories are described as being less complex, less extensive (more limited in scope), more concrete (less abstract), and more comprehensible. This perhaps confounds too many different dimensions to be useful for characterising theories; more careful authors focus on the level of generality dimension, as do Davey *et al.*, below. Attention is now being focussed on ‘middle-range’ or ‘mid-level’ theory at the cutting edge of the (quasi-)experimental development impact evaluation literature. The two terms are used interchangeably, but I will adopt ‘middle-range’ as it is the original term employed by Merton and seems to have narrowly more support in the literature. Defined by their level of generality, middle-range theories are contrasted not just with grand, general theories, but also with intervention-specific theories of change. Davey *et al.* (2018, p.2) describe ‘middle-range’ theories as follows:

> *They are more general than the specific theories of change that describe how the inputs associated with a particular intervention will lead to its intended outcomes, but they are not as high-level as grand theories such as Marx’s theory of class stratification or Foucault’s theory of governmentality, which will offer very limited analytic traction for informing specific interventions.*

The Davey et al. paper is one of the ‘inception papers’ published by the Centre of Excellence for Development Impact and Learning (CEDIL). CEDIL was recently set up with funding from DFID to ‘develop and test innovative approaches to impact evaluation and evidence synthesis in low-income countries’ (CEDIL, 2019a, footer). One of the three programmes of work under which CEDIL will fund a range of projects with budgets up to 1m GBP is ‘Generalising evidence through middle range theory.’ Projects that seek funding submitted proposals in the second half of 2019 and began work in late 2020 or early 2021. That a large emerging area of cutting-edge research should be explicitly cast in terms of ‘middle-range’ or ‘mid-level’ theory is a huge opportunity for this research project to be perceived as relevant and to be useful.

The CEDIL call for proposals says that middle-range theories should ‘explain processes, mechanisms and behaviours in sufficiently general terms to find application in multiple contexts and circumstances’ (CEDIL, 2019b). The first CEDIL working paper, *To Boldly Go Where No Evaluation Has Gone Before: The CEDIL Evaluation Agenda*, describes realist CMO configurations as ‘an operationalisation of mid-level theory’ (Masset and White, 2019, p.12). In their usage, a mid-level theory ‘explains how a program works in a plurality of contexts, and therefore implicitly generalises.’ The generalised programme theories developed in Chapters Five and Six of this thesis are not programme-level CIMOs, explaining only ‘how a particular intervention will lead to its outcomes,’ rather, they are mid-level theories that seek to explain how a type of intervention can be expected to work (or not) across a variety of contexts. Deriving lists of MICCs to generate reporting requirements for evaluations that wish to facilitate the transferability of their results is one example of how middle-range theory can be used to improve evaluation practice. CEDIL’s interest in funding work ‘to elaborate how to develop and use mid-level theories,’ and the broader discursive swing towards ‘theory-based evaluation’, is a sign of the usefulness of this research project to methodologically interested evaluators. Chapter Nine builds an answer to the primary research question in these terms in order to maximise the usefulness of this research to evaluators in international development.

9 Towards a systematic account of (quasi-)experimental impact evaluation quality, and further insights for the evaluation literature

In this chapter, the primary research question is answered. To build an answer to the primary research question, answers to each research subquestion must have been given. Chapter Seven developed an answer to part a) of the operationalised interpretation of research subquestion one, arguing that the novel use of realist programme theory mapping to generate lists of MICCs to assess the transferability of evaluations in both sets was a success. Chapter Eight developed an answer to research subquestion two. Accordingly, it falls to the first two sections of this chapter to develop an answer to the second part of research subquestion one. This answer lays the groundwork for an answer to the primary research question, which is developed using the concepts and the terminology that Chapter Eight has argued maximise its chances of being useful. This answer is given in Section Three of this chapter, arguing that method choice is not systematically associated with quality, neither in theory nor in practice. To avoid giving the impression that 'anything goes' it is demonstrated that analysis of programme theory provides a systematic guide to method choice and to the appraisal of existing evaluations. It is further argued that there are multiple ways of generating programme theories that can be used in these ways – and that the realist method used in this thesis is only one of them. Section Four argues for two secondary advantages of adopting a theory-based approach to evaluation design; these are for internal validity and for the practice of evidence synthesis. Section Five argues that embracing Realism is not required to generate middle-range programme theory and MICCs.

Section Six discusses some ways in which this research could be extended in future work. The research questions and the location of their answers are recapped in the box below for convenience.

Research question recap

Research question	**Answered in**
Research subquestions:	
1.	
a) Can realist programme theory mapping be adapted to create a tool to assess the transferability of (quasi-)experimental development impact evaluation results?	Chapter 7
b) If so, what can it tell us about the systematic differences, if any, between (quasi-)experimental impact evaluation methods regarding the extent to which they report on the barriers and enablers of intervention mechanisms present in the study context and the extent to which they report the degree to which different mechanisms are responsible for changes in outcomes…	
iii. as they are currently used?	Chapter 9
and	
iv. as they might be used?	Chapter 9
2. For epistemic communities of development experts, what are the shared notions of validity concerning what counts as a 'high quality' (quasi-)experimental impact evaluation? Further, what are the features of these accounts that are valued by members of the community, and what unresolved puzzles or nascent crises undermine them?	Chapter 8
Primary research question:	
Can we give a useful, systematic account of the relative merits of evidence generated using different (quasi-)experimental development impact evaluation methods that goes beyond internal validity to also consider the extent to which methods facilitate the transfer of results to other contexts? If so, how?	Chapter 9

9.1 WHAT HAS BEEN LEARNED ABOUT (QUASI-)EXPERIMENTAL IMPACT EVALUATION METHODS AS THEY ARE CURRENTLY USED?

Chapter Seven argued for the utility of the novel method employed in Chapters Five and Six by demonstrating that much had been learned about each of the two cases through the employment of the method. This section interprets both cases to draw tentative lessons about the wider evaluation literature. As Chapter Four, Subsection 4.1.4 has described, the two cases were selected on the basis that they were the best-studied intervention-outcome pairs that could be identified. The evaluations contained in the set of evaluations belonging to each case are not a random sample of the total population of (quasi-)experimental impact evaluations of development interventions. Therefore, it is not possible to construct a mathematical argument for the representativity of the sample. However, I do make a more limited theoretical argument that the sample may be usefully informative of features of the population. This argument can be made based on the features of the evaluations studied that might be expected to be causally relevant to the methodological decisions made in their design.

We might reasonably expect that evaluators from different disciplines design their evaluations differently. This could be a source of bias that undermines any more general lessons drawn from the two cases studied. However, the evaluations in the two sets studied are from either side of the disciplinary divide between social sciences and health that runs more-or-less down the middle of the evaluation of development interventions. As Chapter Four, Subsection 4.1.3 mentioned, 53% of (quasi-)experimental development impact evaluations to date were published in social science journals, as working papers, or as reports (Sabet and Brown, 2018). The remainder were published in health journals. As working papers are much more likely to be published by social scientists and economists, and as reports represent a small fraction of evaluations published, Sabet and Brown (2018) estimate that roughly 50% of evaluations of development interventions are published by social scientists and economists and the remainder are published by public health researchers and epidemiologists. Although no statistical argument can be made for the representativity of the specific disciplinary backgrounds of authors of evaluations in the two sets studied, we can be confident that authors are drawn from a wide variety of disciplinary backgrounds and from both the major fields in (quasi-)experimental development impact evaluation.

It might also be expected that the type of programme evaluated might systematically affect the design of evaluations, introducing bias to any more general lessons drawn from the two cases studied. On this score, too, the evaluations in the sets can be argued to be broadly representative of the wider literature. This is because public health interventions, education interventions and social protection interventions represent 65% of all interventions evaluated in the (quasi-)experimental development impact evaluation literature (Sabet and Brown, 2018). The first case is built around evaluations of an intervention that is described as belonging in the social protection and education topic areas in the 3ie repository of (quasi-)experimental impact evaluations. The second case is built around evaluations of a public health intervention. Therefore, lessons originating from these cases are not likely to be unrepresentative as a result of the broad classes of interventions studied. These arguments for some representativity are intended to be cautious. As with the disciplinary background of the authors, more fine-grained analysis of the topic areas studied in the total population of evaluations would reveal ways in which the two cases examined are not representative. However, on the basis of the information available, these two cases are a reasonable basis on which to build what Pawson (2000, p.300) calls a ‘representative cases argument.’ In the following subsections, three lessons from the data will be presented that together constitute an answer to research subquestion 1 b) i). That is to say, they are an answer to the question ‘what can the cases tell us about the systematic differences, if any, between (quasi-)experimental impact evaluation methods, as they are currently used, regarding the extent to which they report on the barriers and enablers of intervention mechanisms present in the study context and the extent to which they report the degree to which different mechanisms are responsible for changes in outcomes?’ Section 9.2 looks at what has been learned about different (quasi-)experimental impact evaluation methods as they might ideally be used, rather than how they were actually used by the evaluators behind the sets of evaluations.

9.1.1 In general, contextual markers are under-reported

The first lesson that can be drawn from the evaluations examined about (quasi-)experimental development impact evaluations as they are currently conducted is that contextual markers are under-reported. Chapter Seven, Subsection 7.1.2 has discussed this in depth. Figures 7.7 and 7.13 make this very clear for both Case One and Case Two, respectively:

Figure 7.7: Case One group scores reported as proportion of maximum possible score using box plot with individual observations plotted

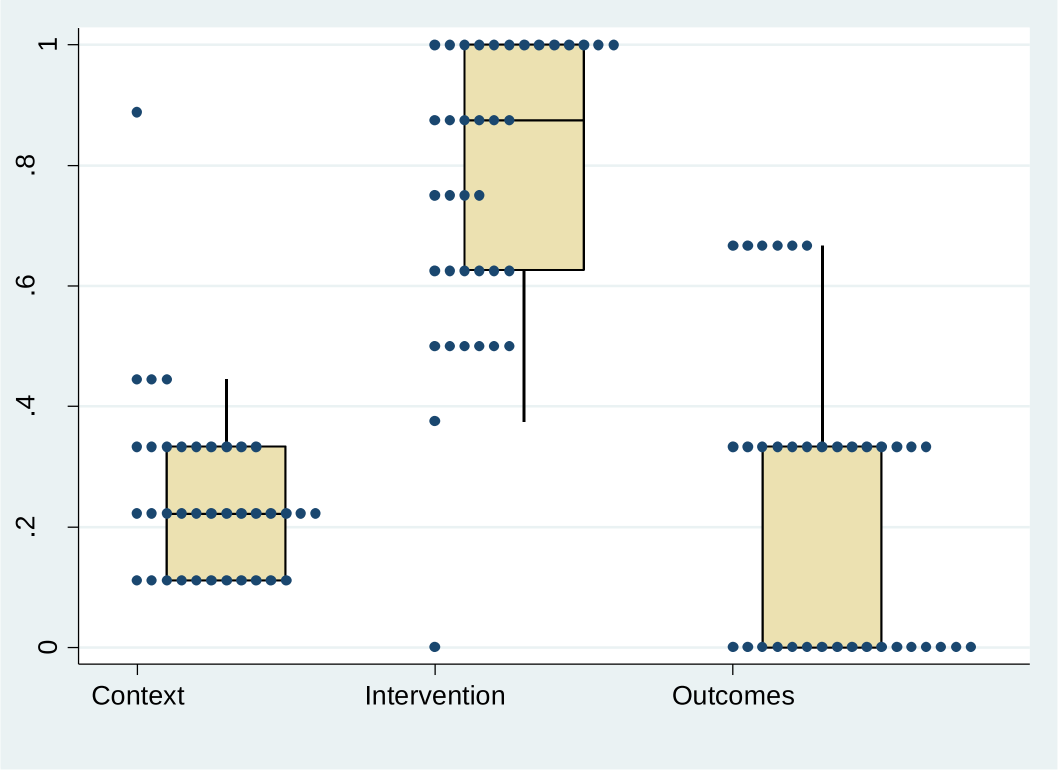
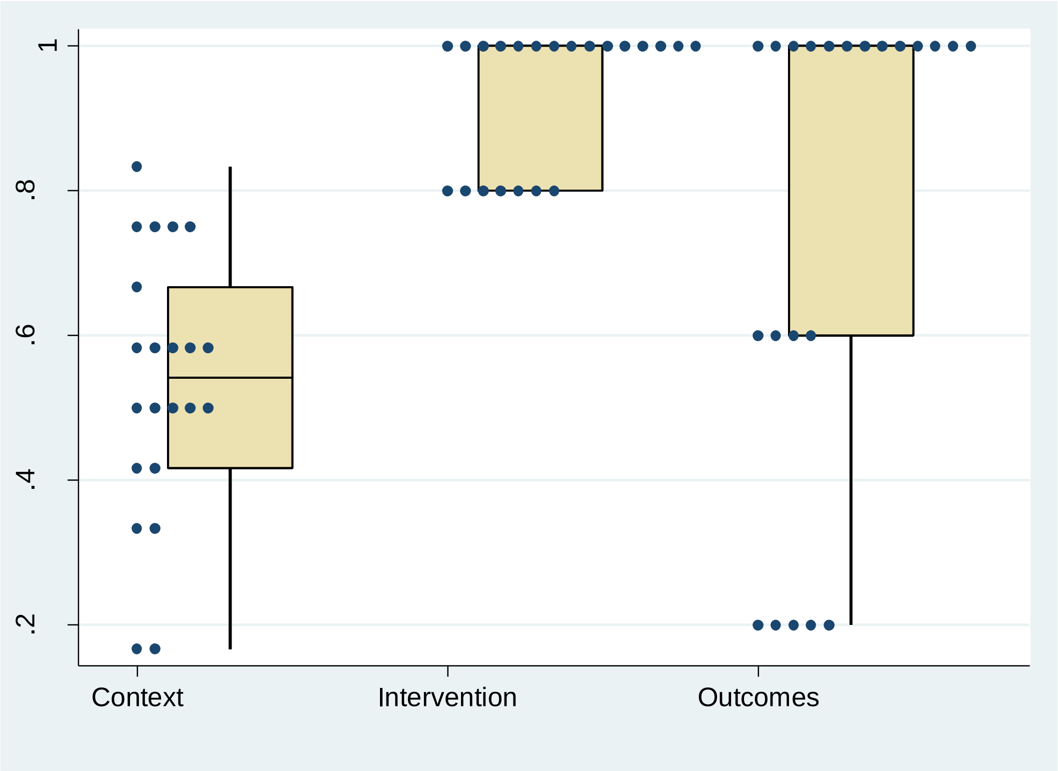


Figure 7.13: Case Two group scores reported as proportion of maximum possible score using box plot with individual observations plotted



As Subsections 7.1.7 and 7.2.6 of Chapter Seven make clear, the fact that some markers of context are almost never reported for evaluations in the sets critically undermines the evidence base in both cases. For example only 5% of the evaluations of the effect of deworming interventions on child weight report child footwear use in the contexts in which they are conducted. This is surprising and unfortunate because transmission of soil-transmitted helminths (STHs) from bare earth to bare foot is universally acknowledged to be a key vector in STH infection. It is therefore possible that all or most of the variation in effectiveness of deworming interventions for child weight is due to different levels of use of footwear in the communities studied. Although this is not likely, it cannot be ruled out. That such a glaring omission has gone unremarked and uncorrected in the literature is an inditement of current and past practice in the (quasi-)experimental impact evaluation of development interventions. A focus on accrediting interventions as effective or not to the exclusion of basic thinking about intervention theory has impoverished the evidence base in the study of development. However, as Chapter Eight, Section Three has noted, this is now widely recognised as a problem.

As the figures above also make clear, intervention implementation is better engaged with and reported on. This is a familiar pattern for those who have examined theory-based evaluation in many fields over many years. Weiss (1997, p.48) makes a distinction between 'programme theory,' which includes the moderating effect of context on outcomes, and 'implementation theory,' which concerns features of the implementation intervention that affect outcomes. She goes on to argue that it is very common for purportedly theory-based evaluations to merely examine differences in outcomes in terms of what she terms 'implementation variables.' This pattern, observed by Weiss in sociology, public health and management, also appears to characterise the (quasi-)experimental development impact evaluation literature. Indeed, the good reporting of markers of implementation in Case One is what allows García and Saavedra (2013) to construct a mathematical model for the effect of certain features of CCT implementation on school enrolment and other outcomes and to test that model through meta-regression.

9.1.2 Intermediate outcomes are better reported than outcomes disaggregated by subgroups affected by different mechanisms

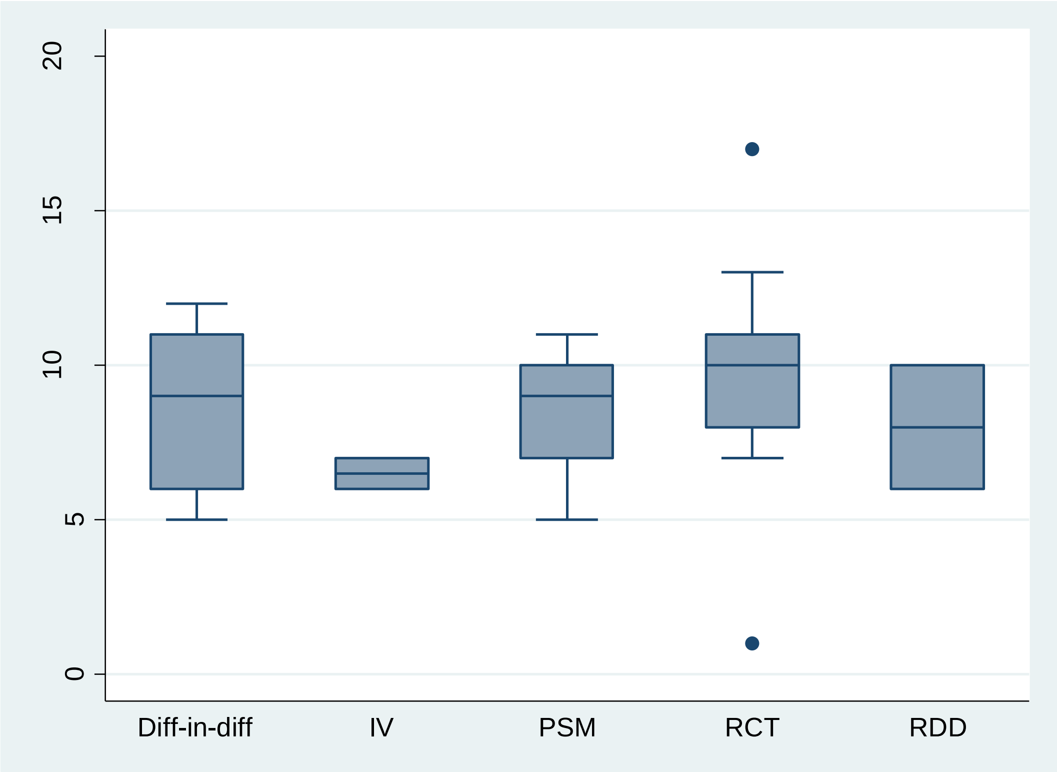
One arresting difference between Figures 7.7 and 7.13, above, is the extent to which outcome markers were reported by evaluations from the two different cases. In case two, outcome markers, on average, were reported by 75% of evaluations in the set. For outcome markers in case one, this proportion was 40%. Although both sets of markers were outcomes reporting data that were required to distinguish between the actions of different intervention mechanisms, they were of different types. In case two, the data were intermediate outcomes. In case one, they were disaggregations of primary outcome data by subgroups that programme theory predicted would be differently affected by different mechanisms. This reflects a difference in the importance placed on each type of outcome reporting in the methodological literature. The importance of intermediate outcomes is widely discussed in the theory-based evaluation literature, and this discussion may be affecting evaluation practice (Glasgow and Linnan, 2008; Rogers, 2007; White, 2018). The importance of reporting outcomes data disaggregated by subgroups that theory tells us might be differently effected by the intervention is not widely discussed, though many talented evaluators are led to report primary outcomes measures in this way, especially when the evaluation has been explicitly designed to test theory, as in Akresh et al. (2012a). A notable exception to this last observation is disaggregation by gender, of course.

High-level social theory tells us that all-pervasive gendered expectations and structural discrimination means different genders will be differently affected by social interventions. This is so widely accepted that gender-disaggregated reporting of outcomes has become institutionalised as best practice in all social science. The increasingly widespread acceptance of the importance of theory-based evaluation provides fertile ground for realists to insist that this practice is also extended to middle-range theory. This would mean that outcomes data are systematically reported disaggregated over all causally-critical features of different subgroups in the study population, as they already are by Akresh et al. (2013) and other talented evaluators.

9.1.3 Presence/absence of markers is not correlated with method choice

As Chapter Seven, Subsection 7.1.2 has discussed in detail, total marker score is not related to method choice for evaluations in Case One. Figure 7.2, below, demonstrates this visually. This exercise was not possible for evaluations in Case Two, which, typically of the public health literature, did not display the same methodological diversity. All the evaluations identified for the second case were RCTs.

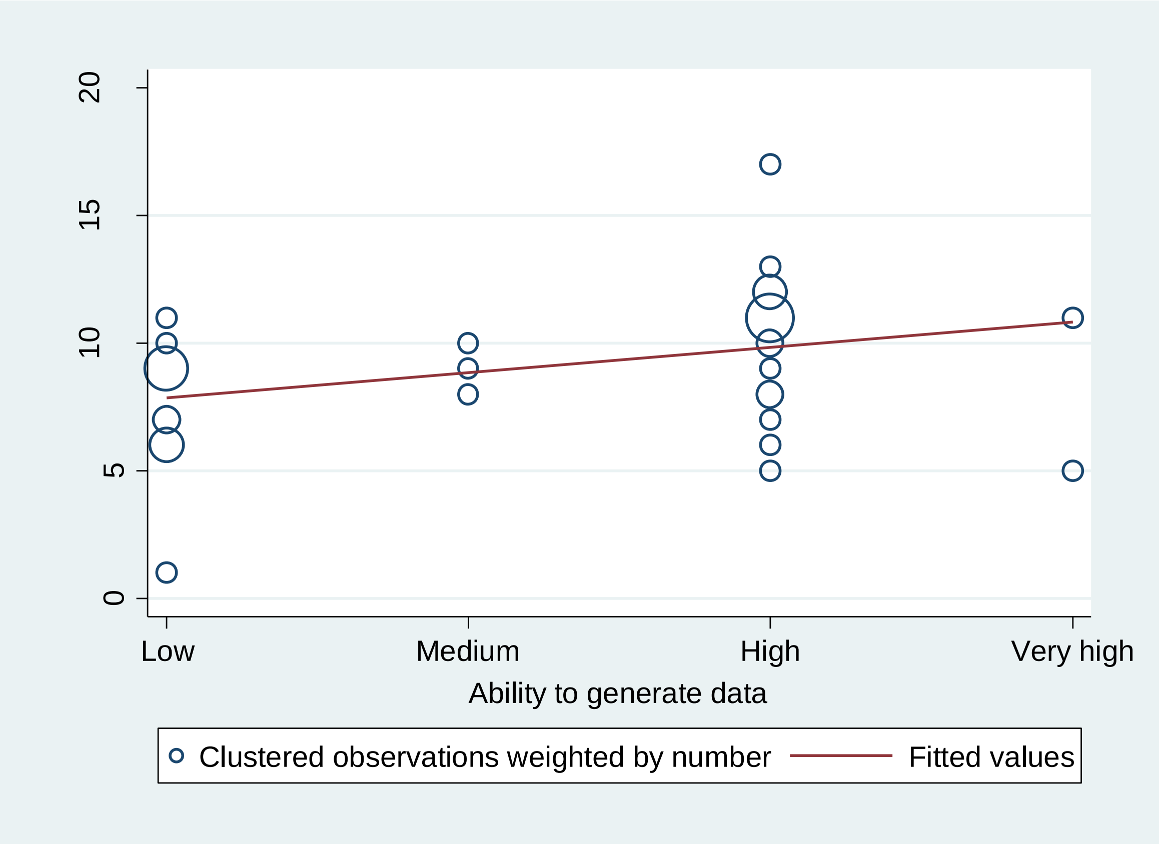
Figure 7.2: Distribution of total marker scores disaggregated by method for evaluations in Case One



In Figure 7.2, method choice is treated as a nominal variable, one that has no inherent order. However, as Chapter Seven, Section 7.1.2 discusses, using the Maryland Scientific Methods Scale allows for the ordering of methods according to a widespread perception of their ability to facilitate internal validity. Ordering methods in this way revealed a very weak positive correlation between total marker score and Maryland Scientific Methods Scale score for the method employed, suggesting that higher perceived propensity to facilitate internal validity was in fact positively correlated with propensity to facilitate the transferability of results. However, this correlation was weak and had an over 11% chance of occurring by chance, were there in fact no relation between the two variables. Therefore, the hypothesis that evaluation methods' ability to facilitate internal validity is negatively correlated with transferability of results, as is widely believed, has no support in the observations from Case One. Rather, the inverse has some weak statistical support. However, the conclusion that for evaluation methods ability to facilitate internal validity is *positively* correlated with transferability of results is only very weakly supported statistically, and we have no theory as to why this might be the case to provide support for this claim.

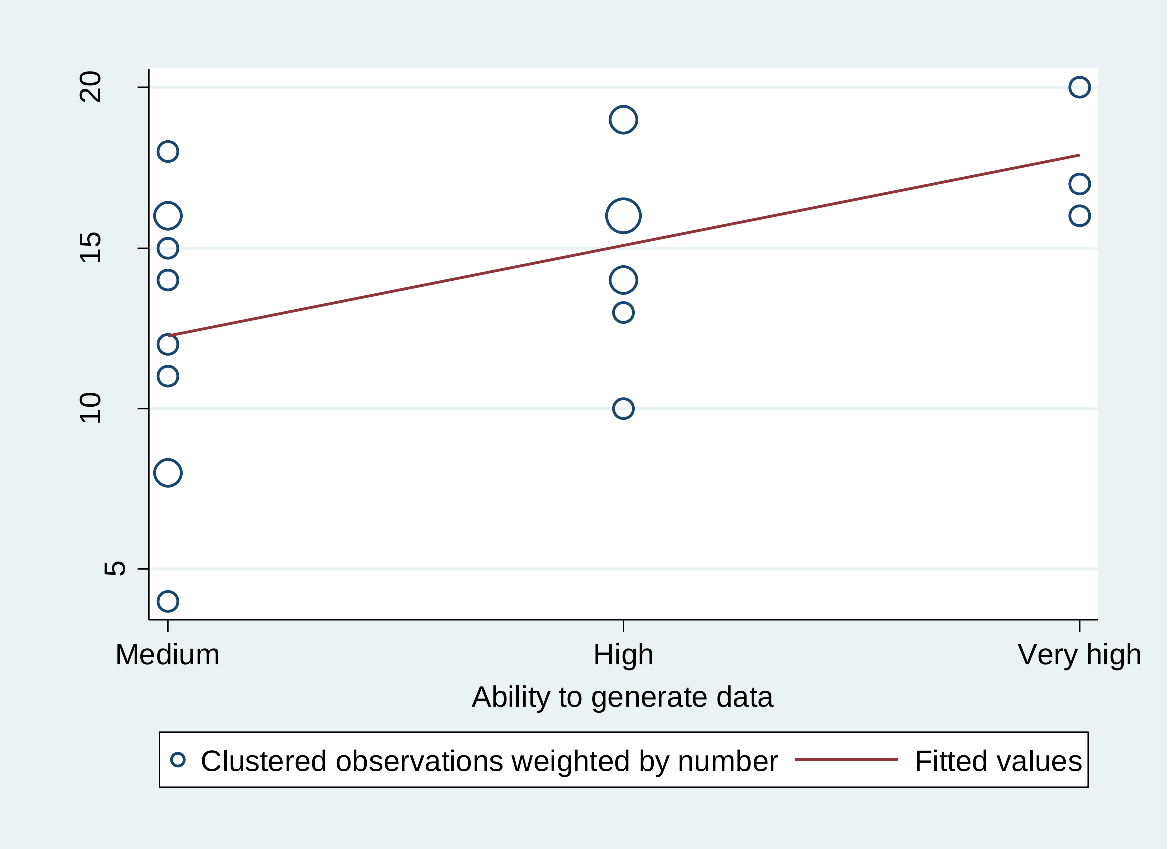
A better explanation of differences in total marker score between evaluations is provided by investigating the relationship between ability to generate data and total marker score. For both sets of evaluations, the ‘ability to generate data’ variable from the corresponding dataset was coded as an ordinal variable taking a value of one to four, with one corresponding to ‘Low’ and four corresponding to ‘Very High’. Calculating Spearman’s Rank Correlation Coefficient for the association between this variable and total marker score for evaluations in Case One yielded a coefficient of 0.3758 with a p-value of 0.0219. The coefficient corresponds to a low to moderate level of correlation. The p-value represents a 2.19% chance that such a finding would arise by chance, were there in fact no association at all. This is a much stronger level of statistical support than the preceding correlational analysis. Figure 7.5 communicates this information visually:

Figure 7.5: Association between total marker score and ability to generate data for evaluations in Case One



Conducting the same exercise for evaluations in Case Two yields a coefficient of 0.5059 and a p-value of 0.0163. This represents a stronger correlation that would arise by chance 1.63% of the time, were there in fact no association between the variables. Figure 7.11 communicates this information visually:

Figure 7.11: Association between total marker score and ability to generate data for evaluations in Case Two



On their own, these observed correlations are not sufficient evidence of a causal connection between total marker score and ability to generate data. However, we can combine these observations with compelling theory. As described in Section Four of Chapters Five and Six, evaluations that employ custom surveys at baseline and endline were scored ‘High’ or ‘Very high’ on ability to generate data. Such evaluations should be expected to be better able to generate and report data corresponding to the various contextual features and disaggregations of outcome data that represent 60% (12 of 20) of all the causal markers for evaluations in Case One. Evaluations that scored ‘Low’, by contrast, were based on secondary reinterpretation of administrative data that would be less likely to permit the generation and reporting of these data. Evaluations that scored ‘Medium’ employed custom surveys only at baseline. Combining the correlations observed with this theory, the inference to the best explanation is that that total marker score, which constitutes the level of facilitation of transferability, is being driven by the use of well-designed custom surveys rather than by any particular identification strategy. There is little relationship, then, contingently expressed in the data, between method choice and ability of an evaluation to facilitate transferability. Subsection 9.2.3 of this chapter discusses whether any relation might hold necessarily between ideally implemented methods of evaluation. In this way, the supposed trade-off between internal and external validity is investigated.

9.2 WHAT HAS BEEN LEARNED ABOUT (QUASI-)EXPERIMENTAL IMPACT EVALUATION METHODS AS THEY COULD BE USED?

The previous section has highlighted three key findings that can be tentatively generalised to the (quasi-)experimental development impact evaluation literature based on the investigation of the two cases identified in Chapter Four. These findings constitute the key pieces of information that have been learned about (quasi-)experimental impact evaluations as they are currently conducted. However, to investigate the possibility of building a systematic account of (quasi-)experimental development impact evaluation quality, it is also important to draw out any lessons about the practice of (quasi-)experimental development impact evaluation as it might be better conducted. To this end, this section presents four findings.

9.2.1 Many studies could do much better without much change in budget.

It might be claimed that the demand for better reporting of markers of intervention causation implies an increase in the cost of evaluation to such an extent that it is not a realistic demand. This subsection argues that that is not the case. As described in Section Four of Chapters Five and Six, evaluations in both sets were coded for their ability to generate data relating to the markers of intervention causation. Evaluations that employed custom surveys at baseline and endline were coded as ‘High’ for this variable. Evaluations which also included some additional data collection, for example through a mixed methods approach incorporating focus grouping or key stakeholder interviews were coded as ‘Very High.’ In Case One 14 of 37 evaluations were coded as ‘High’ or ‘Very High’. In Case Two this figure was 12 of 22. Combining across the two cases, the figure is 26 of 59 or 44%.

The evaluation with the highest marker score in Case One scored 17 of 20 and was coded ‘High’ for ability to generate data. For Case Two, the highest-scoring evaluation scored 20 of 22 markers. However, this evaluation was coded ‘Very High’ as it made use of qualitative methods to explore the health practices of recipients in the sample. Two evaluations in Case Two reported 19 of 22 markers while only being scored ‘High’ for their ability to generate data. Every marker for both cases was reported by at least one evaluation that scored ‘High’ for ability to generate data, except for the one marker in Case One that was reported by no evaluations. This represents an enormous unmet potential of evaluations across both cases to report markers of intervention causation at little or no additional cost. Data for every marker across the two cases could have been generated by the custom baseline and endline surveys that were employed by 44% of evaluations. All markers bar one actually were reported by some such evaluation. This implies no increase in cost for 44% of evaluations to achieve a perfect marker score, had those evaluations interrogated the programme theory that underpinned the intervention and used it as a guide for what data to generate and report relating to contextual features, to implementation features and to intermediate outcomes and reporting of outcomes for subgroups.

56% of evaluations were constrained in their ability to report markers of intervention causation by relying on secondary data analysis, for example the use of administrative data, rather than being able to employ custom surveys. For these evaluations it is much more challenging to estimate how much reporting was constrained by the available of data and how much it might have been improved through more attention to programme theory. However, the imperfect reporting of markers relating to simple features of intervention implementation suggests that at least some of the failures to report markers are not driven by lack of availability of data. Information relating to these makers, such as the implementing institution behind the programme, were readily available to all researchers.

9.2.2 High quality (quasi-)experimental impact evaluations include activities currently considered a part of process evaluation

The UK Medical Research Council guidelines for process evaluation describe the generation of quantitative data relating to mediator and moderator variables as well as intermediate outcomes as parts of process evaluation (Moore et al., 2015, pp.26, 55). However, considering the demands of transferability makes it clear that these data are not optional add-ons that may or may not be collected by a (quasi-)experimental impact evaluation. If the (quasi-)experimental impact evaluation aims to generate insights that are useful beyond the study population at the time of the study, then these data must be generated and reported. One of the authors of the MRC guidelines for process evaluation agrees in private correspondence, saying that when (quasi-)experimental impact evaluation is done properly, the quantitative element of process evaluation that the guidelines call for need not exist as it has already been incorporated into the (quasi-)experimental impact evaluation.

This is also reflected in the quote from Participant Seven in which they say, emphasis added, ‘there are studies where we can’t use the outputs because they are so poorly reported. It can be everything from not describing the context, so that will include the context of the problem, to describing in detail what is the intervention. This was the design, what actually happened, *what did your process evaluation suggest happened*, and you know… just basic things like sample size, means and standard deviations.’ If reporting what the process evaluation said is critical information that determined the utility of a (quasi-)experimental impact evaluation to a systematic reviewer, then such information should not be considered optional to a well-conducted (quasi-)experimental impact evaluation. It may be the case that the line between process evaluation and (quasi-)experimental impact evaluation needs to shift to include the quantitative estimation of intermediate outcomes within (quasi-)experimental impact evaluation, or perhaps the line between impact and process evaluation will dissolve entirely as the theory-based approach to evaluation takes hold. After all, the use of ‘mixed-methods’ in impact evaluation of complex interventions has been considered ‘required’ by the UK’s MRC for the last twenty years (Campbell et al., 2000, p.1).

9.2.3 There is no connection between method choice and the facilitation of transferability

Subsection 9.1.3 of this chapter has shown that, for evaluation in both cases and by extension in all likelihood the wider development evaluation literature, the best explanation of different levels of facilitation of transferability is the ability to generate data rather than the (quasi-)experimental impact evaluation method chosen. No contingent connection has been observed between method choice and transferability in the (quasi-)experimental development impact evaluation literature. However, there may be some necessary, *a priori*, connection between (quasi-)experimental development impact evaluation methods as ideally conducted. Although randomised controlled trials tend to rely on the creation of custom baseline and endline surveys and therefore tend to be associated with a higher ability to generate data, this is not a necessary feature of the method. Many of the early evaluations of PROGRESSA were based on randomised allocation of units of assignment to the programme, but the evaluators did not have full control over the administrative data collected. Likewise, for observational studies, although evaluators do not have control over assignment to treatment they may conduct custom surveys of recipients and non-recipients in order to collect data that is not already being generated, for example by local government.

It is often asserted that there is a necessary connection between method choice and ‘external validity’. Specifically, that there is a trade-off between ‘external validity’ and ‘internal validity’. By this is meant that researchers choosing a method to address their research questions must choose between methods that better facilitate internal validity, and those that better facilitate external validity. Chapter Three, Section One has argued that ‘external validity’ is ambiguous and has framed this research project in terms of the ability of methods to facilitate transferability. That is, the ability of methods to provide premises for valid arguments for the extent to which treatment effects can be expected to hold in some other context. In Chapter Four, it was argued that the reporting of the markers of intervention causation as derived from programme theory constitutes the provision of sufficient premises for an argument for transferability. If this is the case, then there is no necessary connection between the facilitation of transferability and (quasi-)experimental impact evaluation method choice.

The existence of the supposed trade-off between internal validity and external validity for (quasi-)experimental impact evaluations relies on a specific reading of the meaning of 'external validity.' That is, a reading that limits its scope to an assessment of the representativity of the study population as regards some larger total population of possible recipients of an intervention. This intended meaning of external validity was labelled 'generalisability' in Chapter Three, Section One. The trade-off exists in this case because it is assumed that the creation of a compelling counterfactual group of non-recipients that are otherwise similar to recipients is only possible within certain subsets of the total population of interest, and more exacting standards of internal validity restrict the possible subsets more than do less exacting standards. This is a very specific situation, however. Generally even evaluations of pilot projects intended to be 'scaled up' to a large population are not interpreted based on a naïve generalisation from sample to population that does not take account of characteristics of the pilot population. As soon as these characteristics are being taken account of and the projection of treatment effects in the sample to the population is being based on the causally relevant characteristics of each, the relevant questions are not questions of generalisability but of transferability as defined in Chapter Three.

9.3 ANSWERING THE PRIMARY RESEARCH QUESTION

> *Can we give a useful, systematic account of the relative merits of evidence generated using different (quasi-)experimental development impact evaluation methods that goes beyond internal validity to also consider the extent to which methods facilitate the transfer of results to other contexts? If so, how?*

We are now in a position to answer the primary research question. We can give a useful, systematic account of the relative merits of evidence generated using different (quasi-)experimental development impact evaluation methods that goes beyond internal validity to also consider the extent to which methods facilitate the transfer of results to other contexts. This useful, systematic account is laid out in this section and based on centring a middle-range theory of intervention causation across contexts. Generating and analysing such a theory provides premises for an argument for method choice for new evaluations and also premises for arguments for the transferability of existing evaluation evidence to new contexts.

In the first subsection of this section I argue that creating a realist mapping of programme theory to generate MICCs and using it to guide data collection and therefore report all relevant MICCs has been shown to be necessary and sufficient for facilitating arguments for the transferability of findings. I further argue that all (quasi-)experimental impact evaluation methods are equally compatible with this approach. In later subsections I argue that making middle-range programme theory central to evaluation design and interpretation in the way demonstrated in this research project can put flesh on the bones of 'theory-based evaluation' as it relates specifically to (quasi-)experimental impact evaluations. I also argue that it can provide a systematic guide to method choice for new evaluations, and to the interpretation of existing evaluations. This account does not constitute a hierarchy of methods, but it improves on 'it depends' by trying to provide practical advice for evaluation practitioners and users. Building on the insights generated in Chapter Eight through the consultation of participants and the literature, I argue that it is likely that many evaluators will be receptive to this advice. I further argue that the capital R Realist ontology and epistemological strategy used in this research project are not necessary to ground the generation of programme theory and the derivation of MICCs. I show that alternative methods can be used, referring to work I have published with Nancy Cartwright and co-authors.

First, a reminder of the scope of this research project. There are many purposes for which evaluations are conducted and the relevant criteria upon which to judge evaluation quality differ by purpose. For example: evaluations may be conceived as accountability mechanisms whose function is only to demonstrate the level of effectiveness of a specific programme (Cutt and Murray, 2000). In this case, no argument for the transferability of results will be necessary. A high quality (quasi-)experimental impact evaluation in this instance is one that produces the desired estimates of the treatment effect.[70] In Chapter Two, this research project was framed as an attempt to discuss evaluation quality as it relates to the attempt to make better policy and design better programmes. Therefore, for the rest of this section it will be assumed that the purpose of (quasi-)experimental impact evaluation is to contribute to knowledge of the world in general rather than just one specific context at one time. In practice, evaluations are conducted for a variety of reasons, and the arguments of this section will hold true to the extent that the preceding assumption is met.

9.3.1 Method choice does not determine transferability

As subsection 9.2.3 has argued, all (quasi-)experimental impact evaluation methods are equally capable of generating transferable insights. In order to do this, they must report the markers of intervention causation in context (MICCs) that are implied by a programme theory of that class of intervention that they are evaluating an example of. The ability to do this is not related to (quasi-)experimental impact evaluation method choice, but rather to the evaluators' ability to generate data, and the extent to which they use programme theory to determine the data generated and reported. Subsection 9.2.2 has argued that this may mean measuring and reporting parameters that have traditionally been considered the preserve of process evaluation, such as intermediate outcomes and the moderating factors that we have termed barriers and enablers of intervention causation in context. As Subsection 9.2.1 argued, this is not an impossible demand. Many (quasi-)experimental impact evaluations could meet this requirement at very little or no additional cost. Chapters Five and Six show that realist programme theory mapping can be adapted to generate an aggregated programme theory for a type of intervention from which MICCs can be derived. Chapter Seven demonstrated that the resulting insights were informative, generating insights for both the literatures from which evaluations in the sets were

[70] This does not necessarily mean an estimate of the average treatment effect on the treated, of course. Many features of the distribution may need to be estimated in order for the needs of accountability to be satisfied. For example, a strongly positive average treatment effect with negative effects for a significant minority may be unacceptable. Further, it may be important to estimate treatment effects over an entire population rather than an experimental subgroup. Further, high variance in treatment effects may mean that higher bias is an acceptable cost for a larger sample size (Ravallion, 2018, pp.6–7). These considerations and others mean that a randomised controlled trial, though it may be the best design in some circumstances, is not the 'gold standard' even for an evaluation designed to achieve such context-specific goals.

drawn. The first section of this chapter demonstrated that assessing and comparing the reporting of MICCs by the evaluations in both sets generates informative insights for the development evaluation literature as a whole. Further, as discussed in Chapter Eight, evaluations in the first case are increasingly likely to report MICCs. This could not be demonstrated for evaluations in the second set, but this is in large part because earlier evaluations in the second set were already reporting a much higher proportion of MICCs than were evaluations in the first set, making an increasing trend harder to detect. So MICCs must be measured and reported to facilitate arguments for transferability, and this is a realistic demand for evaluations using all methods that has been demonstrated to be possible for two different types of interventions.

It is a striking finding of this research that the supposed trade-off between internal validity and external validity relies on a specific reading of the meaning of 'external validity.' That is, a reading that limits its scope to an assessment of the representativity of the study population as regards some larger total population of possible recipients of an intervention. In fact, in practice, for evaluations in the sets, because RCTs tended to be associated with a higher ability to generate data, those experiments were likely to have better facilitated arguments for the transferability of their findings than were quasi-experimental or matching evaluations. In both sets, the evaluation that has the highest MICC score is an RCT.[71] It might be argued that this leaves us 'back where we started,' so to speak, in that (quasi-)experimental impact evaluation methods have been found not to generate more or less transferable insights than each other either in theory or in practice.[72] If some methods are more able to generate more reliably internally valid estimates than others, then this forces us to re-embrace a uni-dimensional account of evidence quality based on internal validity. However, this is not the case. In the following sections I argue that generating a middle-range programme theory of the sort of

[71] Benhassine *et al.* (2015) for the first case and Zhang *et al.* (2017) in the second case.

[72] Though in practice evaluation methods that can rely on analysis of secondary data to generate internally valid estimates of treatment effects, such as matching methods, may lose some of their comparative cost and convenience advantage over experiments (though not their precision advantage in some cases), or accept less readily transferable findings. This is because they may have to use new data generation instruments to generate the MICCs required for transferable findings.

intervention concerned can provide the basis for a systematic approach to method choice and to the transfer of results between contexts.

9.3.2 Choosing an appropriate method

The primary research question asks about the ‘merits’ of different (quasi-)experimental impact evaluation methods for development interventions, considering internal validity and transferability. There are two key junctures at which such considerations are relevant for practitioners and for users of evaluation. The first is at the point of method choice when tasked with conducting a counterfactual, effectiveness evaluation of an intervention with many treatment units (a large *N*). This subsection sketches a systematic approach to choosing a method in this situation. If one is tasked with evaluating such a programme, a deliberately uncharitable model of the *randomistas’* likely advice would be that one should randomise allocation to treatment if at all possible, and if not, one should fall back on the closest possible quasi-experimental method to randomisation, resorting to a matching approach only if necessary. In this way, one will be able to generate a minimally biased estimate of the treatment effect, and this is the most important consideration. This is likely not the view that any of the RCT authors would endorse; it is too simplistic. However, it is the view that has been understood from their work by some, and they sometimes appear very close to endorsing it (Banerjee, 2006; Duflo, 2017; Imbens, 2010).

Many RCT authors in the literature have sought to add nuance to this view. Ravallion (2018, p.6) points to ethical concerns with using random allocation to treatment in situations where a) one could afford to treat all potential recipients or b) one has prior theory that suggests which beneficiaries will benefit the most, and randomisation therefore constitutes a harm to those recipients. Ravallion’s point b) is an instance of a more general observation: to the extent that our evaluation is intended to increase our knowledge about the world, our existing theory, and our confidence in that theory should determine our design choices when specifying an evaluation. This is what Participant Two referred to as a ‘questions-driven’ approach to evaluation, and what other Participants referred to as a ‘theory-based’ approach.

As Chapter Eight has argued, the theory-based approach to evaluation is an emerging hegemony in the evaluation of development interventions, but it is unclear what practical consequences this has for (quasi-)experimental impact evaluation design and interpretation. This research project has shown the great power of a realist research strategy to put flesh on the bones of a theory-based approach to (quasi-)experimental impact evaluations. Chapters Five and Six showed that realist programme theory mapping can be used to generate theories that can be analysed to determine which markers of intervention causation must be reported to facilitate transferability. Chapter Seven showed that analysis of the reporting of those markers by evaluations of a given intervention-outcome pair could identify gaps in the literature to guide research priorities. The success of these efforts can be abstracted from to provide a guide for method choice when designing an evaluation.

As every programme is based on a theory of how the intervention will combine with context to change outcomes, that theory should be elicited from programme designers in as much detail as possible before making any evaluation design decisions. Then the theory can be assessed to determine the design that will lead to the most learning by focussing resources on the most uncertain parts of the model. In an extreme case like that suggested by Ravallion's b), we might have solid theory in which we have high confidence that subgroup A of recipients are the most critical group to which to extend intervention T, and have just enough budget to extend the intervention to all members of A. In this case, the only ethical assignment would be to assign all of A to treatment.[73] If A can be broken down into clusters, and if it is possible, affordable, and ethical to observe participants for a period of no intervention before the intervention begins, then a stepped-wedge, cluster-randomised design may be preferable (Hemming et al., 2015). If

[73] For example, consider an 'ultra-poor graduation programme' funded from a regional budget and well resourced (Banerjee et al., 2015). Targeting suggests that we could fund the intervention for all of the ultra-poor in the region. Alternatively, we could include as many households again that are above the asset index cutoff used for targeting, randomise allocation of households to control or intervention in this larger population, and see what happens. If we have high confidence in the efficacy for the ultra-poor and are unsure of the effect for slightly better-off households, then we might plausibly be ethically obliged to target only and all of the ultra poor, depending on the specifics of the situation.

these conditions are not met, the best approach may be to develop a quasi-experimental design to compare members of A with the best counterfactual available.

Alternatively, our solid theory might suggest that T could be extended to either subgroup A or subgroup B, and we do not have the available resources to treat more than half of either subgroup. In such a case, if randomisation is possible, it is very unlikely not to be good value for money compared to a perhaps very marginally cheaper quasi-experiment. However, our theory suggests extremely high variance in treatment effects. In this case, we should assign resources so as to increase precision as much as possible. This might mean eschewing a factorial trial designed to test different forms of the intervention or the effects of different potentially moderating characteristics of units in favour of a simple two-armed trial.

Alternatively again, our theory of the action of the intervention in context might well be very uncertain, a more plausible scenario in development. In that case random assignment to treatment might allow us to reduce bias from the very many unknown confounding factors active in our specific context, and we could look at our programme theory and determine which aspects of it could most be improved by learning in our context. Perhaps there are two possible mechanisms for action of the intervention that have always been tested concurrently. In this case we can use a cluster-randomised design to evaluate two different variants of the programme (if we have the budget to maintain sufficient power to detect likely effects) in order to separate those mechanisms and test their relative importance in our context. There are many excellent examples of this type of design in the literature, such as Baird et al. (2011).

The unifying theme behind these three examples of method and design choice is that programme theory and our certainty in it, as applied to our specific context, can determine design decisions. The process of assessing the transferability of evaluations in both cases has demonstrated the utility of a realist approach to (quasi-)experimental impact evaluation. By applying this approach to evaluation method and design decisions in this subsection I have demonstrated that, for this use case, we can give a systematic, useful account of the relative merits of evidence generated using different (quasi-)experimental impact evaluation methods that goes beyond internal validity to also consider the extent to which methods facilitate the transfer of results to other contexts.

9.3.3 Transferring treatment effects to some given context

There is another use case for an account of method quality for the evaluation of development interventions that must be addressed in answer to the primary research question. This is the use of evidence issued from different methods evaluating different study interventions to inform predictions of likely effects in a target context. As has been argued in the previous section, there is no systematic link between (quasi-)experimental impact evaluation method and the transferability of results. However, this does not mean that the realist approach to (quasi-)experimental impact evaluation championed in this thesis can offer no guidance when assessing the available impact evidence and attempting to make predictions. In this case, no lexicographic preference for any given (quasi-)experimental impact evaluation method is suggested. The best that one can do is to look at treatment effects from other contexts in the light of the model of intervention causation in which one has the highest confidence and attempt to translate results between contexts. Results can be translated by looking at the markers of intervention causation that are reported for interventions and comparing them to data on those same markers in the target context. Treatment effects from different evaluations should be discounted to the extent that they are imprecise or likely to be biased. It is important to remember that imprecision is just as valid a concern as bias. Ravallion (2018, p.6) points out that in case of high variance in outcomes, a more accurate estimate can be expected to result from a biased but more precise estimate than from an unbiased but imprecise estimate. To see why, imagine two normal distributions of treatment estimates, one with higher bias and low precision, and one with no bias and lower precision. If the difference in precision is sufficiently high compared to the level of bias, randomly drawing from the biased distribution will result in an answer closer to the true value more of the time than drawing from the unbiased distribution. This is why, for example, a large matching study might be more informative evidence than a smaller RCT.

In addition, treatment effects from different source evaluations should be modified by the extent to which causally relevant characteristics of study context and study intervention implementation are different from observations for those same markers in the target context. This is very close to some existing approaches such as the approach suggested by Bates and Glennerster (2017). They suggest a four-step approach to transferring results to a target context:

> *Step 1: What is the disaggregated theory behind the program?*
>
> *Step 2: Do the local conditions hold for that theory to apply?*
>
> *Step 3: How strong is the evidence for the required general behavioral change?*
>
> *Step 4: What is the evidence that the implementation process can be carried out well?*
>
> (*ibid*)

The approach suggested in this thesis is more general, in that intervention mechanisms are not all assumed to target 'behavior change.' It is also more precise in that it provides a description of the form that the theory behind the programme must take in order to be useful, and provides an exhaustive approach to identifying all of the relevant MICCs that must be considered to assess Bates and Glennerster's steps two and four.

There are also similarities between the approach suggested in this section and that of Bonell et al. (2006). Those authors suggest that facilitation of transferability (generalisability in their usage) requires that (quasi-)experimental impact evaluations:

- *Include process evaluations as integral elements*
- *Develop evidence based theories about how intervention processes are influenced by context and how processes might differ if interventions are implemented in other sites*
- *Report the extent to which their participants are representative of the population being targeted*

- *Describe the prevalence of the needs being met by the intervention, informed by clear hypotheses about the intervention's mechanism.*

(*ibid*, p.348).

Further, these reporting requirements should ensure that the (quasi-)experimental impact evaluation empirically assess the 'acceptability' of the intervention in context, the 'feasibility of delivery', the extent of 'coverage' of the target population achieved, and the extent to which the intervention responded to 'local needs.' Clearly for Bonell et al. transferability can be assessed in the way suggested in this Chapter, by assessing the similarities between the MICCs present in study and target contexts, with those MICCs being based on programme theory. The emphasis on particular types of MICC is a result of Bonell and co-authors' situation in the public health literature, but the account is recognisably similar to the more general account developed here.

Incomplete reporting of MICCs presents a barrier to the approach to transferability suggested in this thesis as well as similar approaches. As Chapter Seven demonstrated for each of the two cases studied, currently even evaluations of the best-studied interventions do not report all MICCs. This means that current attempts to transfer results are frustrated by incomplete data, and are limited in the insights they can generate. As subsection 9.1.1 has shown, the situation is better for intervention-design-related markers than for markers of contextual features. It is already possible, in the cash transfers literature and for deworming interventions, to draw conclusions from the evidence available about the ways in which intervention design decisions are likely to effect outcomes. For example, García and Saavedra (2017) are able to conduct a meta-analysis in contexts so-far studied of the effects of intervention design decisions represented by intervention markers on outcomes for conditional cash transfer programmes for school enrolment. However, those same authors are not able to say anything about the effects of characteristics of context represented by contextual markers, as these markers are not reported by enough of the studies in their sample.

So, an attempt to follow the systematic method above will run into incomplete data, in practice. However, this should not be seen as a weakness of the method. It is better to know that one is acting on the basis of incomplete information, and to use what information is available, than to boldly act unaware of one's ignorance. It is currently possible to appraise the existing deworming or CCT evidence from other settings and to use the method above to attempt to construct a reasoned argument for the extent of transferability of findings to a target setting, using what MICCs have been reported. This attempt would have to acknowledge the gaps in information available, particularly the fact that more information about intervention design in existing settings is available than contextual information about the causally relevant features of context in those settings. However, it would be a superior attempt than one based on a naïve extrapolation of the average treatment effect across existing settings unadjusted by an attempt to compare study and target contexts. It would also be superior to an attempt to adjust for some features of contexts when interpreting results from existing evaluations, but which did not use programme theory to ensure that that attempt considered all and only the MICCs of relevance.

9.3.4 The need for rules of combination in realist (quasi-)experimental impact evaluation design and interpretation

In both of the previous sections, the uses of programme theories that are suggested require rules of combination for the moderating effects of the markers of intervention causation on outcomes in order to make quantitative arguments. The programme theories developed in Chapters Five and Six did not contain such rules because they were not present in the literature from which those theories were synthesised. This is a limit of current evaluation practice. Mathematical models in papers tend not to encompass features of context or many features of intervention implementation, if they are specified at all. Such models are inadequate to underpin the prediction of treatment effects in new target contexts, therefore, and must be extended to include all markers of intervention causation implied by programme theories in order for the approach suggested in this chapter to produce quantified estimates. This may seem like an extremely onerous demand to place upon evaluators. However, as Cartwright (2008) points out, such models are already being used implicitly by anybody attempting to make policy recommendations on the basis of the existing literature. The demand made here is merely to bring these models into the open and state them explicitly. It will be necessary for authors in many literatures to confront the fact that these models are simplistic and are not well supported by empirical evidence. That is the inevitable consequence of treating transferability as secondary to internal validity, but it might be overcome. As Cartwright (2008, p.41) puts it: ‘It’s no good ducking the problem. We’d better just get on with figuring out how to make this all as simple and user friendly as possible.’

The realist device of the CMO or CIMO configuration is an essentially qualitative device that needs to be translated into a mathematical model in order to permit quantitative analysis. This is not to say that expressing programme theory in realist terms is not necessary. Expressing theory in realist terms forces us to construct a theory that pays proper attention to the interactions between intervention, context, mechanism and outcome.[74] However, once the theory has been expressed in this qualitative way it must be translated into a model that admits of mathematical specification in order to facilitate quantitative analysis and prediction, if that is desired.

One way of translating qualitative theory into a model that can have mathematical rules of combination attached is the use of directed acyclic graphs (DAGs) (Pearl and Mackenzie, 2018). This approach is increasingly common in medicine but has not yet been adopted in economics (Imbens, 2019). The adoption of some such translation technique is necessary for realist programme theory to be given quantitative predictive power, and it is desirable to further clarify the relationships between elements of the theory, which are sometimes seen to be confused by the CMO presentation (Hawkins, 2014). In new work funded by CEDIL and begun in February this year (2021) I am working with co-authors on the development of a process we have called POInT, Process-Outcome Integration with Theory. This work builds on Pearl and Mackenzie’s (2018) use of DAGs as well as Humphries and Jacobs’ (2015) work on Bayesian integration of qualitative and quantitative data, with those latter two authors being co-authors on this project. We have partnered with existing FCDO-funded evaluations of development interventions, and

[74] Section 9.6 argues that there are also non-realist alternatives.

are eliciting the programme designers' middle-range theory of intervention causation across contexts. We will use this theory to generate MICCs that should be included in data collection efforts across the (therefore more integrated) process and (quasi-)experimental impact evaluation. In addition, we will represent this theory as a DAG in order to permit the quantitative analysis of the interaction between nodes in the DAG based on the results of the evaluation. We will also work with programme designers to attach prior probabilities to the relationships in the DAG and show how confidence in those relationships can be transparently, reliably updated based on evaluation data.

Another way of translating qualitative theory into a model that can have rules of combination attached is the use of set-theoretic or 'configurational' approaches such as qualitative comparative analysis (QCA). QCA was first developed by Ragin (1987) and has been much developed since, with several different flavours of QCA now established (Befani, 2016). QCA is a method for comparing between cases to identify any combinations of characteristics (or conditions) of those cases that are either necessary, sufficient or an INUS condition for a given outcome. This analysis is either conducted strictly (in 'crisp-set QCA') on combinations of binary conditions with a binary outcome, or using fuzzy-set logic as in fuzzy-set QCA, which allows for four or six-valued conditions and outcomes. Multi-Value QCA goes even further in this direction, allowing more values (*ibid*). QCA is typically used to compare conditions between a small or medium number of cases. However, it is possible to use QCA on larger datasets (Fiss, Sharapov and Cronqvist, 2013).

Clearly, there is a natural fit of QCA analysis with cases described in terms of lists of MICCs. If particular values for MICCs are reinterpreted as conditions, and re-rendered in binary (or quaternary, sexternary etc. in the case of fuzzy-set or multi-value QCA), then they may be suitable for QCA analysis to organise values of MICCs into configurations of conditions with causal effects on outcomes of interest. This sort of analysis could be conducted at two different levels. Firstly, it would be possible for an evidence synthesiser to analyse several (quasi-)experimental evaluations of an intervention-outcome pair in different contexts and assess the extent to which QCA could suggest any configurations of values of MICCs that were necessary, sufficient or formed INUS conditions for the successful action of the intervention. Alternatively, consider the case of a single (quasi-)experimental impact evaluation which had been informed by a generative theory of intervention causation and had reported all of the MICCs in the way suggested in this thesis. This (quasi-)experimental evaluation could be extremely fruitfully combined with QCA by taking each treatment unit as a case (including those in control groups) and using a QCA approach to this large *N*. This might uncover configurations of MICCs with causal effects at the treatment unit level. This sort of large-*N* analysis could even be conducted across multiple evaluations to create a very large *N*.

9.3.5 Caveats to scope and originality

It is worth noting that the primary research question seeks to extend an account of method quality beyond external validity *only* to include the transfer of results to other contexts. There are, of course, other dimensions to impact evaluation quality such as construct validity, transparency and relevance to goals (Stern et al., 2012; Stern, 2015). It is outside of the scope of this thesis to develop a general account of impact evaluation quality which responds to all these quality dimensions. Attempts to do so have been made by multi-author working groups as reported by Stern et al. (*ibid*) and Leeuw and Vaessen (2009). Rather, I more modestly demonstrate the feasibility and utility of considering transferability when assessing methods *in the case that* a (quasi-)experimental approach to the impact evaluation has been deemed the most appropriate approach. I hope that by developing a specific framework for incorporating generative accounts of intervention theory into (quasi-)experimental impact evaluations, I have provided an analysis that is deeper, albeit more narrow, than the analysis in those more general treatments of impact evaluation quality.

Besides pointing out the necessarily limited scope of this project, I would like to place some caveats on its originality, in case it appears that I am over-claiming. This thesis argues that it is necessary for (quasi-)experimental impact evaluations to identify, measure, and report MICCs in order to produce insights that are useable beyond the study context. In one sense, this is a repetition of the old claim that ‘mixed-methods’ are ‘required’ in the evaluation of social interventions (Campbell et al., 2000, p.1). General accounts of how methods should be mixed and under what circumstances have been produced since at least 1989 (Greene, Caracelli and Graham, 1989). The wide-ranging general accounts of impact evaluation cited in the previous paragraph also have much to say about the utility of mixing methods and about how it should be done. The originality of this thesis consists (only) in providing a concrete account of how realist programme theory mapping can be 1) combined with all (quasi-)experimental impact evaluation methods to produce more transferable insights and 2) used as a method for assessing the transferability of insights arising from such methods when conducing a review of evidence.

9.4 ADDITIONAL BENEFITS OF MAKING MIDDLE-RANGE PROGRAMME THEORY CENTRAL TO EVALUATION

The previous section has argued that an approach to (quasi-)experimental impact evaluation centred around middle-range programme theory provides a useful, systematic guide to method choice and the application of (quasi-)experimental impact evaluation results from study context to target context. This section argues that this approach could have two additional benefits.

9.4.1 Improving internal validity

Matching prior to randomisation on causally relevant observables improves statistical power, allowing for more precise estimates of treatment effects using smaller sample sizes (Raudenbush, Martinez and Spybrook, 2007). The markers of intervention causation in context that must be reported to facilitate transferability constitute such causally relevant observables. Therefore, producing an explicit programme theory complete with a list of the markers of intervention causation in context would not just benefit transferability, but also internal validity, even for RCTs. Further, what is observable and what in ‘unobservable’ is not fixed. Especially when conducting randomised evaluations, careful thought to what we chose to observe can increase both internal validity and transferability. With good theory, backed up by ANCOVA, we can match and assess balance on causally relevant observables rather than the somewhat arbitrary lists of socio-economic characterises upon which balance is generally assessed, many of which are often not supported as causally important by programme theory.

9.4.2 Improving meta-analysis

Banerjee and Duflo (2008, p.16) defend an experimental approach to (quasi-)experimental impact evaluation by saying '[i]f we were prepared to carry out enough experiments in varied enough locations, we could learn as much as we want to know about the distribution of the treatment effects across sites conditional on any given set of covariates.' There are two problems with this defence. One is that it is not possible to carry out an experiment in all contexts, maybe not even in most of the contexts where we would like to examine the effects of development interventions. In particular, infra-marginal impacts of programmes, rather than the roll-out of those programmes to previously unserved areas is only possible using observational designs (Ogden, 2016, p.59).

The second problem with Banerjee and Duflo's defence in so far as it was intended to apply to their avowedly successionist approach, is that their level of scepticism about programme theory does not allow them a strategy for discovering which are the relevant 'covariates.' A realist approach to evaluation design based on the need to develop and test middle-range theory of the mechanisms of intervention causation solves this problem. Once programme theories are specified in a way that permits the derivation of the markers of intervention causation in context, the list of crucial 'covariates' can be consistently derived and improved upon by authors in the literature. The more evaluations report a given marker, the greater will be the power of meta-analysis to quantify the relationships between contextual features, features of intervention implementation and the degree of effectiveness of specific intervention mechanisms. Subsection 8.4.2 in the previous chapter identified a demand for systematic, actionable improvements to current evidence synthesis techniques. This demand can be satisfied by a realist approach to specifying and testing middle-range theory in the form of context-intervention-mechanism-outcome configurations and mathematical models specifying their combination for particular intervention designs.

9.5 EMBRACING REALISM IS NOT REQUIRED TO GENERATE MIDDLE-RANGE PROGRAMME THEORY AND MICCS

The preceding sections might sound like a demand that all impact evaluators of development interventions sign up to Pawson's brand of realism. As Chapter Eight, Subsection 8.5.2 acknowledged, this is not a realistic demand. This is because many evaluators perceive this brand of realism, capital R Realism if you will, to be set up in opposition to experiments, with Realist Evaluation often spoken of as an alternative rather than a complement to (quasi-)experimental impact evaluation. It is true that this thesis has made use of a realist research strategy. I believe it has been a useful framework capable of underpinning the work done. However, in order to derive the MICCs for a given intervention and to realise the broader advantages of centralising a middle-range programme theory, it is not necessary to make use of CMO or CIMO configurations and all of the other machinery of Pawson's Realism. There are other ways of generating programme theory that are also adequate. This is because the key ontological move required, from a successionist understanding of causation to a generative understanding of causation is more fundamental than is the choice of a realist research strategy.

In a CEDIL methods paper published with co-authors including Nancy Cartwright, my co-authors and I set out to describe the features that a programme theory of change (pToC) must have in order to facilitate reliable predictions of programme success in new contexts (Cartwright et al., 2020). We build on existing work in diverse fields including some Realist work, but the resulting framework is not recognisably Realist. We do not even use the word 'mechanism' in order to avoid confusion for example between Jon Elster's (2015) use of the term and Ray Pawson's (1997). Rather, we demonstrate how cross-context, general theories of change for a type of intervention can be built by breaking down intervention causation into individual causal steps underpinned by a single tendency principle. This tendency principle implies the barriers and enablers (or support factors and derailers) that might act on each step of the causal change, as well as the 'safeguards' that can act to prevent derailers from derailing. In that paper, we develop three examples of middle-range programme theories of change for three sorts of intervention, one of which is conditional cash transfers to increase school enrolment. It may interest readers to compare the account of programme theory given in terms of CIMO configurations in Chapter Five of this thesis with the account in terms of tendency principles in Section 5.4 and Figure 16 of Cartwright et al. (2020). Despite not being grounded in realism, the pToC framework is a way of generating middle-range theories of intervention causation across contexts that permit the derivation of MICCs and are sufficient for the other purposes described earlier in this section.

Similarly Iversen and Lanthorn (2016) develop a non-Realist account of how to build a theory of change sufficient to ensure that evaluations generate and report the data required to facilitate an argument for the transferability of their findings. They build on Cartwright and Hardie (2012), Woolcock (2013) and Hotz et al. (2005), synthesising their approaches to facilitating transferability as well as adding some of their own requirements to plug the gaps they identify in these existing frameworks. The result is an account somewhat similar to Cartwright et al. (2020), though the 'tendency principles' which link nodes in the theory of change and their implied 'barriers, enablers and safeguards' of that account are absent. Instead, Iversen and Lanthorn (2016) create lists of 'assumptions' associated with each node on the step-by-step theory of change, and distinguish between 'critical' and 'substantive' assumptions, where critical assumptions are necessary to the acting of the mechanism that connects two nodes in the theory of change, whereas substantive assumptions can merely amplify or diminish effects. The result is less comprehensive than Cartwright et al. (2020) or the CIMO method presented here. This is because the lists of 'assumptions' have to compress a discussion of mechanism (or tendency principle), barriers, enablers, and safeguards into a single sentence. This risks critical details being overlooked when formulating the assumptions and therefore being absent from the theory of change. Nevertheless, Iversen and Lanthorn's framework is another example of a non-realist method for generating middle-range theories of change that apply to programmes across settings.

What these examples demonstrate is that realism is not a necessary ingredient in constructing middle-range theories of change capable of underpinning lists of MICCs and the other types of analysis described earlier in this section. Rather, the critical ingredient is an acceptance of the legitimacy and importance of a generative account of intervention causation in context. Chapter Eight has argued that the ‘theory-based approach’ to evaluation is an emerging hegemony in development evaluation, with practitioners endorsing the importance of this approach in writing and in the interviews conducted for this research project. If this endorsement means anything at all, then it is an endorsement of generative accounts of causation. Clearly a successionist approach, which holds that causes are not sensible objects of study, is not compatible with an attempt to derive theories of how and why interventions work and then use those theories to guide evaluation. Therefore, the evaluation experts that this research seeks to influence have already crossed the crucial ontological bridge. What remains is to put flesh on the bones of a theory-based approach to (quasi-)experimental evaluation by showing how and why middle-range theories of intervention causation across contexts can guide method choice and arguments for the transferability of results between contexts. I have attempted to do so in this thesis.

The answer to the primary research question, then, is that we can give a useful, systematic account of the relative merits of evidence generated using different (quasi-)experimental development impact evaluation methods that goes beyond internal validity to also consider the extent to which methods facilitate the transfer of results to other contexts. This useful, systematic account was laid out in this chapter and based on centring a middle-range theory of intervention causation across contexts. Generating and analysing such a theory provides premises for an argument for method choice for new evaluations and also premises for arguments for the transferability of existing evaluation evidence to new contexts.

9.6 FURTHER POTENTIAL AVENUES FOR FUTURE RESEARCH

Subsection 9.3.4 has discussed further work I am conducting with colleagues on the POInT project and also suggested that QCA might be used to create rules of combination for particular values of MICCs in some cases. This section describes some further potential avenues for future research. In the examination of the two cases studied in response to research subquestion one, I have mostly restricted myself to assessing whether evaluations reported MICCs or not. I have also asked and answered questions about why some evaluations reported a higher proportion of MICCs than others, and what the costs and benefits of reporting a higher proportion of MICCs would be. It has not been possible, while keeping this research project tractable, to address two other important questions which might fruitfully be studied.

9.6.1 Were MICCs *used* by evaluations in the set to investigate questions of transferability?

Throughout this thesis I have made occasional mention of excellent evaluations within the studied cases that not only reported a high proportion of MICCs but were designed to learn about the moderating effects of those MICCs on the effect of the intervention. For example in the CCT case, Baird et al. (2011) was designed to test the moderating effect of conditionality and Akresh et al. (2013) was designed to test differential effects on more or less marginal children. Subsection 5.2.3.1.6 discussed at length the fact that Benhassine et al. (2013) was designed to test for the existence of a new mechanism, which I have called the ‘labelling mechanism’ and which has only more recently been added to the consensus understanding of how CCTs function.

Future work could investigate for all evaluations in both sets what analyses were conducted that explicitly attempted to use MICCs to test middle-range theory of the functioning of intervention mechanisms, as well as the methodological strengths and limitations of those analyses. This work was not possible within the scope of this research project. However, my strong prior is that it would provide confirmatory evidence of the usefulness and importance of conducting evaluations based on generative middle-range theories of intervention causation. In addition, I believe that it would demonstrate for specific evaluations that reporting of MICCs can provide sufficient premises for a convincing argument for the transferability of findings. Benhassine et al. (*ibid*), in particular, could be used as interesting specific case study of the power of reporting MICCs because it reported 17 of 19 markers. This could be contrasted with an evaluation that reported a low proportion of MICCs to demonstrate concretely what effect this had on the transferability of findings.

9.6.2 What does the evidence from each set of evaluations say about the scope of application of the intervention?

An even more ambitious piece of future research could perhaps extend the sort of analysis mentioned at the end of the previous subsection to the entirety of each case. Having assessed the arguments in favour of transferable findings that are warranted by each evaluation in each case, it might be possible to say something useful about the sum of such arguments for the whole case. In other words, it might be possible to say for which sorts of contexts and which ways of implementing the intervention good evidence exists for the effectiveness of the intervention, for which contexts and ways of implementing the intervention is not likely to work, and where there are knowledge gaps. Of course, such an assessment would not be an assessment of the total evidence available. Both cases are limited to (quasi-)experimental impact evaluations of the studied intervention-outcome pair, so this line of research would only be able to summarise the transferable evidence *arising from (quasi-)experimental impact evaluation*. To inform decisions in any target context, these insights would need to be combined with insights arising from different methods such as observational studies of the target context, and perhaps small-n piloting evaluations using a process-tracing or other non-counterfactual impact evaluation approach. Despite the risk of being over-interpreted as a summary of '*the* available evidence,' such a project could be valuable. This thesis has identified some glaring knowledge gaps for example in the significance of footwear-prevalence for the effectiveness of deworming interventions targeting child weight. However, the further work suggested in this subsection would be able to go beyond such insights to potentially identify more subtle, more specific knowledge gaps like those that exist for particular types of contexts and/or ways of implementing.

10 Conclusion

Seeking to influence public policy and to make it more evidence-based has long been an objective of social scientists. Perhaps this is especially true of those who study ‘development.’ We have chosen to specialise in an area that is the study of change, whether we presuppose that this change is for the better or not. For many, the idea that academic research does contribute to public policy change that does improve the lot of the poor and marginalised is the motivating factor that drove them to development research. The rise of the *randomistas* in the evaluation of development interventions reflects, for many, an uncritical excess in this manner of thinking (Ogden, 2016). For true believers, this new(ly fashionable) way of learning about development has the potential to transform the relationship between policy and research for the better, ushering in a new age of poverty reduction based on technocratic success.

I began this research project and this thesis from the observation that (quasi-)experimental impact evaluations are a powerful tool and an important cog in the far-from-perfect research to policy machine. It was argued that ‘gold standard’ thinking within (quasi-)experimental impact evaluation, elevating the RCT to the top of a hierarchy of methods, was misplaced. Such a way of thinking is critically undermined by the ‘problem of external validity’, or more precisely, the problem of transferability. This problem rears its head when we try to use (quasi-)experimental impact evaluation results from a given context at a given time to talk about any other context. ‘Gold standard’ thinking has various harmful effects within and beyond (quasi-)experimental impact evaluation, and many have argued that ‘there is no gold standard.’ However, these arguments have done little to change practice. This is perhaps because they sound too much like ‘anything goes’ and impact evaluation practitioners require comprehensible methodological standards to function.

These reflections in Chapter Two led to a candidate primary research question:

> *Can we give a useful, systematic account of the relative merits of evidence generated using different (quasi-)experimental development impact evaluation methods that goes beyond internal validity to also consider external validity?*

In Chapter Three, I argued that 'external validity' was an ambiguous term with at least three major uses or intended meanings in the development impact evaluation literature alone. I accepted that redefining 'external validity' is not a realistic project, but hoped that I might be able to encourage authors to distinguish between three of its meanings by defining them and giving them their own proper terms. This analysis is the first substantive finding of the research project and is being developed into a stand-alone article. I reproduce Table 3.1 here to recap those terms.

Table 3.1: Meanings of 'external validity' with alternative suggested specific terms

	Meaning	**Example**	***ex-ante* / *ex-post***	**Suggested specific term**
A)	The extent to which treatment effects will hold in some other context	'external validity (whether valid inferences are drawn for other projects, either as scaled up versions of that project in the same setting or as similar projects in different settings)' (Ravallion, 2009, p.32)	*ex-ante*	Transferability
B)	The extent to which treatment effects from some sample context will hold in the population context	'external validity—the relevance of the IE [impact evaluation] to the scale-up of the programme in a given country' (Davis et al., 2016, p.65)	*ex-ante*	Generalisability
C)	The extent to which given treatment effects are predictive of the treatment effect across all observed contexts	'the "external validity" of one set of *n* results … *i.e.* how well observed point estimates $Y_1, Y_2, \ldots, Y_n$ can be used to jointly predict the point estimate of another study, Y_j' (Vivalt, 2019, p.12)	*ex-post*	Observed heterogeneity

At the end of Chapter Three, Section 3.1, the primary research question was recast in terms of ‘transferability,’ as below:

> *Can we give a useful, systematic account of the relative merits of evidence generated using different (quasi-)experimental development impact evaluation methods that goes beyond internal validity to also consider the extent to which methods facilitate the transfer of results to other contexts? If so, how?*

This question was split into two component subquestions in order to separate the investigation of 1) the possibility and 2) the utility of the systematic account described. Answering research subquestion one required the development of a novel method capable of assessing the level of facilitation of transferability achieved by (quasi-)experimental impact evaluations employing different methods. As overcoming the problem of external validity requires a generative account of causation, and as realism provides a framework for developing generative programme theories, I appropriate the realist evaluation toolkit to construct this novel method. This method uses tools from realist synthesis as a way of aggregating and synthesising the programme theories behind a group of evaluations, and deriving from them the markers of intervention causation in context (MICCs) that must be reported in order to facilitate transferability. The argument that knowledge of these MICCs for both study and target context constitutes necessary and sufficient premises for an argument for the transferability of (quasi-)experimental impact evaluation findings is the second major finding of this research project and appears in Chapter Three, Subsection 3.2.3.

Chapter Four outlined a method of constructing two sets of evaluations, one for each of two pairs of intervention and outcome. With the help of Vivalt’s AidGrade database, I identified the two best-studied intervention and outcome pairs to serve as my cases; conditional cash transfers for school enrolment, and deworming for child weight (AidGrade, 2013b). Adapting programme theory mapping from realist synthesis allowed me to synthesise the shared accounts of intervention causation underpinning those sets of evaluations and to specify them in realist terms, as context-intervention-mechanism-outcome configurations (CIMOs). From those

CIMOs I derived lists of the contextual features, intervention features and outcomes measures that constituted necessary premises for an argument for the transferability of evaluation results to some other context.

I then assessed evaluations in both sets to examine whether they reported each of the MICCs derived from the theory that underpinned them. This examination permitted the quantitative examination of the level of facilitation of transferability achieved by evaluations in both sets. This process generated a variety of insights for the two literatures from which the two sets of evaluations were drawn, demonstrating the success of the novel method employed. Chief amongst these were the identification of under-reported MICCs which could explain heterogeneity of results between contexts, but which have not been investigated. Whole families of MICCs were reported by very few evaluations in the first set, such as those indicating non-financial barriers to education. In the second set, the standout finding was that footwear prevalence was reported by only one evaluation in the set despite the fact that the absence of footwear is universally acknowledged to be a key determinant of the rate of (re)infection with soil-transmitted helminths. These findings indicate that the method of identifying MICCs and assessing how widely reported they are is not just useful for the purposes of this research project, but also as a means of identifying gaps in evaluation literatures which represent fruitful directions for future research.

In parallel, semi-structured interviews were conducted with a set of experts on the impact evaluation of development interventions, as described in Chapters Four and Eight. These interviews were conducted to determine, in combination with an iterative consultation of the literature, how the account generated in response to the primary research question could be useful. The data generated by these interviews and the accompanying consultation of relevant literature led to an understanding of these development experts as belonging to two closely linked epistemic communities: the *randomistas* and the sceptics. Although these communities were divided on the question of whether RCTs should have a special place in the pantheon of (quasi-)experimental impact evaluation methods, both shared various frustrations with the development impact evaluation literature; chiefly, that transferability was an unsolved problem that current evidence synthesis methods are not able to overcome. In addition, both communities appeared to endorse an emergent paradigm in (quasi-)experimental development impact evaluation, that of theory-based evaluation. Combined with the desire of some on the cutting edge of (quasi-)experimental impact evaluation to put the notion of middle-range theory to use, the ascendence of 'theory-based evaluation,' at least insofar as practitioners near-universally *claim* to practice it, represents an opportunity for the findings of this research project to be warmly received.

Chapter Nine combined the insights from both strands of research to offer an answer to the primary research question. This answer was framed in the terms that Chapter Eight suggested would maximise the chances of its being perceived as useful. At the most basic level the answer to the primary research question is that there is no association between method choice and the facilitation of transferability, either in theory or in practice, at least for the evaluations in both sets. Does this mean that 'there is no gold standard' cannot be improved upon? I argue no.

Firstly, the set of findings around MICCs are an improvement on 'there is no gold standard' by providing practical reporting requirements for impact evaluators seeking to generate transferable results. Chapter Three, Subsection 3.2.3 argues that the investigation and reporting of MICCs is required for a (quasi-)experimental impact evaluation to generate transferable results. This is a useful finding that could guide evaluation practice and would enormously improve the power of evidence synthesis methods including meta-analysis if it were acted upon. Chapter Nine, Subsection 9.2.1 presents an argument that many studies could do enormously better at reporting MICCs and generating more transferable insights at no extra cost. So, reporting the MICCs is not an unrealistic demand on impact evaluators. For evaluations presently relying on analysis of secondary data, some extra cost may be implied if the secondary data are not rich enough to provide the MICCs. If there is no way of generating all MICCs for a study population, then a reduction in the ability of the study to generate transferable insights is implied. Another actionable finding from Chapter Nine is that a requirement to report MICCs may mean that high-quality (quasi-)experimental impact evaluation includes some activities that are currently considered a part of process evaluation. This is not entirely surprising, as the use of 'mixed-methods' in impact evaluation of complex interventions has been considered 'required' by the UK's Medical Research Council for the last twenty years (Campbell et al., 2000, p.1).

More ways in which this research project improves on 'there is no gold standard' are presented in the latter sections of Chapter Nine. I argue that generating middle-range programme theory for an intervention-outcome pair in the way demonstrated in this thesis can provide a guide to method choice for impact evaluators and a guide to transferring effects between contexts. Further, I sketch how tools such as directed acyclic graphs could be used to translate qualitative presentations of programme theory into models that admit of quantitative specification and analysis, where this is required. As Subsection 9.3.4 describes, I have secured funding with co-authors to demonstrate this (amongst other methodological innovations) in partnership with a number of FCDO-funded evaluations of development interventions.

In the final section of Chapter Nine I acknowledge that the preceding sections could be interpreted as a demand for development impact evaluators to 'become realists.' This is a demand that many may baulk at, as Chapter Eight identified. Fortunately, I am able to draw on work published with co-authors including Cartwright (Cartwright et al., 2020), as well as a helpful working paper by Iversen and Lanthorn (2016) to demonstrate that the epistemic strategy and practical tools of realism are not required to generate middle-range theories of intervention causation across contexts sufficient to generate MICCs. This thesis is avowedly realist in the tradition of Pawson (2000). Indeed, I hope it is an example of a work of 'middle-range realism' of the sort that Pawson has called for more of (*ibid*). It is also somewhat critical in the way of Bhaskar and Sayer's realism, with a focus on the issue at hand as a 'social problem of science' (Bhaskar, 2008, p.179; Ravetz, 1995). Nevertheless, I hope that the final section of Chapter Nine demonstrates that it is a pragmatic project, and that there are other epistemic and methodological ladders that can be climbed to get to the same place.

References

Adato, M., Hoddinott, J. and Emmanuel, S., 2010. Combining Quantitative and Qualitative Research Methods for Evaluation of Conditional Cash Transfer Programs in Latin America. In: M. Adato and J. Hoddinott, eds. *Conditional Cash Transfers in Latin America*. Johns Hopkins University Press.pp.26–54.

AidGrade, 2013a. *Coding manual*. [online] AidGrade. Available at: <http://evavivalt.com/wp-content/uploads/2016/05/AidGrade-Coding-Manual.pdf> [Accessed 15 May 2017].

AidGrade, 2013b. *Meta-analysis process*. [online] AidGrade. Available at: <http://www.aidgrade.org/wp-content/uploads/AidGrade-Process-Description.pdf> [Accessed 17 May 2017].

AidGrade, 2013c. *Search terms*. [online] AidGrade. Available at: <http://www.aidgrade.org/wp-content/uploads/AidGrade-Search-Terms.pdf> [Accessed 17 May 2017].

Akoglu, H., 2018. User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, 18(3), pp.91–93. https://doi.org/10.1016/j.tjem.2018.08.001.

Akresh, R., Bagby, E., De Walque, D. and Kazianga, H., 2012a. *Child labor, schooling, and child ability*. The World Bank.

Akresh, R., Bagby, E., de Walque, D. and Kazianga, H., 2012b. Child Ability and Household Human Capital Investment Decisions in Burkina Faso. *Economic Development and Cultural Change*, 61(1), pp.157–186. https://doi.org/10.1086/666953.

Akresh, R., De Walque, D. and Kazianga, H., 2013. *Cash transfers and child schooling: evidence from a randomized evaluation of the role of conditionality*. The World Bank.

Altman, D.G. and Bland, J.M., 1995. Statistics notes: Absence of evidence is not evidence of absence. *BMJ*, 311(7003), p.485. https://doi.org/10.1136/bmj.311.7003.485.

Amarante, V., Ferrando, M. and Vigorito, A., 2013. Teenage School Attendance and Cash Transfers: An Impact Evaluation of PANES. *Economía*, 14(1), pp.61–96.

Angelucci, M., De Giorgi, G., Rangel, M.A. and Rasul, I., 2010. Family networks and school enrolment: Evidence from a randomized social experiment. *Journal of Public Economics*, 94(3), pp.197–221. https://doi.org/10.1016/j.jpubeco.2009.12.002.

Armand, A. and Carneiro, P., 2018. *Impact evaluation of the conditional cash transfer program for secondary school attendance in Macedonia*. 3ie Impact Evaluation Report. International Initiative for Impact Evaluation (3ie).p.48.

Astbury, B. and Leeuw, F.L., 2010. Unpacking Black Boxes: Mechanisms and Theory Building in Evaluation. *American Journal of Evaluation*, 31(3), pp.363–381. https://doi.org/10.1177/1098214010371972.

Atkinson, P. and Coffey, A., 2003. Revisiting the relationship between participant observation and interviewing. In: J.A. Holstein and J.F. Gubrium, eds. *Inside interviewing: new lenses, new concerns*. Thousand Oaks [Calif.]: Sage Publications.

Attanasio, O., Fitzsimons, E., Gomez, A., Gutierrez, M.I., Meghir, C. and Mesnard, A., 2010. Children's schooling and work in the presence of a conditional cash transfer program in rural Colombia. *Economic development and cultural change*, 58(2), pp.181–210.

Awasthi, S., Peto, R., Pande, V.K., Fletcher, R.H., Read, S. and Bundy, D.A.P., 2008. Effects of Deworming on Malnourished Preschool Children in India: An Open-Labelled, Cluster-Randomized Trial. *PLOS Neglected Tropical Diseases*, 2(4), p.e223. https://doi.org/10.1371/journal.pntd.0000223.

Awasthi, S., Peto, R., Read, S., Richards, S.M., Pande, V., Bundy, D., and DEVTA (Deworming and Enhanced Vitamin A) team, 2013. Population deworming every 6 months with albendazole in 1 million pre-school children in North India: DEVTA, a cluster-randomised trial. *Lancet (London, England)*, 381(9876), pp.1478–1486. https://doi.org/10.1016/S0140-6736(12)62126-6.

Baird, S., Chirwa, E., McIntosh, C. and Özler, B., 2010. The short-term impacts of a schooling conditional cash transfer program on the sexual behavior of young women. *Health economics*, 19(S1), pp.55–68.

Baird, S., Ferreira, F.H.G., Özler, B. and Woolcock, M., 2013. *Relative Effectiveness of Conditional and Unconditional Cash Transfers for Schooling Outcomes in Developing Countries: A Systematic Review*. Campbell Systematic Reviews. The Campbell Collaboration.

Baird, S., McIntosh, C. and Özler, B., 2011. Cash or Condition? Evidence from a Cash Transfer Experiment. *The Quarterly Journal of Economics*, 126(4), pp.1709–1753. https://doi.org/10.1093/qje/qjr032.

Banerjee, A., Duflo, E., Goldberg, N., Karlan, D., Osei, R., Parienté, W., Shapiro, J., Thuysbaert, B. and Udry, C., 2015. A multifaceted program causes lasting progress for the very poor: Evidence from six countries. *Science*, 348(6236).

Banerjee, A.V., 2006. Making Aid Work. *Boston Review*. [online] Jul. Available at: <http://bostonreview.net/archives/BR31.4/banerjee.php> [Accessed 5 Sep. 2019].

Banerjee, A.V. and Duflo, E., 2008. *The Experimental Approach to Development Economics*. [Working Paper] National Bureau of Economic Research. Available at: <http://www.nber.org/papers/w14467> [Accessed 31 Jan. 2014].

Barrera-Osorio, F., Bertrand, M., Linden, L.L. and Perez-Calle, F., 2008. *Conditional Cash Transfers in Education Design Features, Peer and Sibling Effects Evidence from a Randomized Experiment in Colombia*. [Working Paper] National Bureau of Economic Research. https://doi.org/10.3386/w13890.

Basu, K., 2013. *The Method of Randomization and the Role of Reasoned Intuition*. Washington D.C.: World Bank.

Bates, M.A. and Glennerster, R., 2017. The Generalizability Puzzle (SSIR). *Stanford Social Innovation Review*, [online] Summer 2017. Available at: <https://ssir.org/articles/entry/the_generalizability_puzzle> [Accessed 10 Apr. 2021].

Becker, H. and Geer, B., 1958. 'Participant Observation and Interviewing': A Rejoinder. *Human Organization*, 17(2), pp.39–40. https://doi.org/10.17730/humo.17.2.mm7q44u54l347521.

Befani, B., 2016. Pathways to change: Evaluating development interventions with Qualitative Comparative Analysis (QCA). *Sztokholm: Expertgruppen för biståndsanalys (the Expert Group* *for Developent Analysis). Pobrane z: http://eba. se/en/pathways-to-change-evaluating-development-interventions-with-qualitative-comparative-analysis-qca*.

Behrman, J.R., Parker, S.W. and Todd, P.E., 2004. Medium-term effects of the Oportunidades program package, including nutrition, on education of rural children age 0-8 in 1997. *Unpublished manuscript*.

Benedetti, F., Ibarrarán, P. and McEwan, P.J., 2016. Do education and health conditions matter in a large cash transfer? Evidence from a Honduran experiment. *Economic Development and Cultural Change*, 64(4), pp.759–793.

Benhassine, N., Devoto, F., Duflo, E., Dupas, P. and Pouliquen, V., 2013. *Turning a Shove into a Nudge? A 'Labelled Cash Transfer' for Education*. NBER Working Paper Series. Cambridge, MA: National Bureau of Economic Research.

Benhassine, N., Devoto, F., Duflo, E., Dupas, P. and Pouliquen, V., 2015. Turning a Shove into a Nudge? A "Labeled Cash Transfer" for Education. *American Economic Journal: Economic Policy*, 7(3), pp.86–125. https://doi.org/10.1257/pol.20130225.

Berndt, C., 2015. Behavioural economics, experimentalism and the marketization of development. *Economy and Society*, 44(4), pp.567–591. https://doi.org/10.1080/03085147.2015.1043794.

Bhaskar, R., 1975. *Realist theory of science ; realist theory of science.* Leeds: Leeds Books.

Bhaskar, R., 2008. *A realist theory of science*. Classical texts in critical realism. London ; New York: Routledge.

Biernacki, P. and Waldorf, D., 1981. Snowball Sampling: Problems and Techniques of Chain Referral Sampling. *Sociological Methods & Research*, 10(2), pp.141–163. https://doi.org/10.1177/004912418101000205.

Blattman, C., 2016. *Why "what works?" is the wrong question: Evaluating ideas not programs*. [online] Chris Blattman. Available at: <https://chrisblattman.com/2016/07/19/14411/> [Accessed 2 Sep. 2019].

Bonell, C., Fletcher, A., Morton, M., Lorenc, T. and Moore, L., 2012. Realist randomised controlled trials: A new approach to evaluating complex public health interventions. *Social Science & Medicine*, 75(12), pp.2299–2306. https://doi.org/10.1016/j.socscimed.2012.08.032.

Bonell, C., Fletcher, A., Morton, M., Lorenc, T. and Moore, L., 2013. Methods don't make assumptions, researchers do: A response to Marchal et al. *Social Science & Medicine*, 94, pp.81–82. https://doi.org/10.1016/j.socscimed.2013.06.026.

Bonell, C., Moore, G., Warren, E. and Moore, L., 2018. Are randomised controlled trials positivist? Reviewing the social science and philosophy literature to assess positivist tendencies of trials of social interventions in public health and health services. *Trials*, 19(1), p.238. https://doi.org/10.1186/s13063-018-2589-4.

Bonell, C., Oakley, A., Hargreaves, J., Strange, V. and Rees, R., 2006. Assessment of generalisability in trials of health interventions: suggested framework and systematic review. *BMJ (Clinical research ed.)*, 333(7563), pp.346–349. https://doi.org/10.1136/bmj.333.7563.346.

Bonell, C., Warren, E., Fletcher, A. and Viner, R., 2016. Realist trials and the testing of context-mechanism-outcome configurations: a response to Van Belle et al. *Trials*, 17(1), p.478. https://doi.org/10.1186/s13063-016-1613-9.

Broadbent, E., 2012. Politics of research-based evidence in African policy debates. *London: Overseas Development Institute*.

Brownson, R.C., Royer, C., Ewing, R. and McBride, T.D., 2006. Researchers and Policymakers: Travellers in Parallel Universes. *American Journal of Preventive Medicine*, 30(2), pp.164–172. https://doi.org/10.1016/j.amepre.2005.10.004.

Burchett, H., Umoquit, M. and Dobrow, M., 2011. How do we know when research from one setting can be useful in another? A review of external validity, applicability and transferability frameworks. *Journal of Health Services Research & Policy*, 16(4), pp.238–244. https://doi.org/10.1258/jhsrp.2011.010124.

Cabinet Office, 1999. *Modernising Government*. Cabinet Office.

Cameron, D.B., Mishra, A. and Brown, A.N., 2016. The growth of impact evaluation for international development: how much have we learned? *Journal of Development Effectiveness*, 8(1), pp.1–21. https://doi.org/10.1080/19439342.2015.1034156.

Campbell, M., Fitzpatrick, R., Haines, A., Kinmonth, A.L., Sandercock, P., Spiegelhalter, D. and Tyrer, P., 2000. Framework for design and evaluation of complex interventions to improve health. *BMJ*, 321(7262), pp.694–696. https://doi.org/10.1136/bmj.321.7262.694.

Card, D. and Krueger, A.B., 1993. *Minimum Wages and Employment: A Case Study of the Fast Food Industry in New Jersey and Pennsylvania*. [Working Paper] National Bureau of Economic Research. https://doi.org/10.3386/w4509.

Carden, F., 2009. *Knowledge to policy: making the most of development research*. Los Angeles : Ottawa: SAGE ; International Development Research Centre.

Cartwright, N., 2007. Are RCTs the Gold Standard? *BioSocieties*, 2(1), pp.11–20. https://doi.org/10.1017/S1745855207005029.

Cartwright, N., 2008. *A Theory of Evidence for Evidence-Based Policy*. Centre for the Philosophy of Natural; and Social Science Contingency and Dissent in Science.

Cartwright, N., 2019. *Evidence for action in new settings: The importance of middle-level theory*.

Cartwright, N., Charlton, L., Juden, M., Munslow, T. and Beadon Williams, R., 2020. *Making predictions of programme success more reliable*. [CEDIL methods working paper:] London, UK: Centre for Excellence in Development Impact and Learning (CEDIL). Available at: <https://cedilprogramme.org/wp-content/uploads/2020/10/CEDIL-methods-WP-N.-Cartwright-Oct-2020.pdf> [Accessed 28 Jan. 2021].

Cartwright, N. and Hardie, J., 2012. *Evidence-based policy: A practical guide to doing it better*. Oxford University Press.

Carvalho, S. and White, H., 2004. Theory-Based Evaluation: The Case of Social Funds. *American Journal of Evaluation*, 25(2), pp.141–160. https://doi.org/10.1177/109821400402500202.

CEDIL, 2019a. About Us CEDIL Centre of Excellence for Development Impact and Learnin. *CEDIL-Centre of Excellence for Development Impact and Learning*. Available at: <https://cedilprogramme.org/cedil/> [Accessed 3 Sep. 2019].

CEDIL, 2019b. *Programme of Work 2: Generalising evidence through middle range theory*. CEDIL Call for Proposals. [online] London, UK: Centre for Excellence in Development Impact and Learning (CEDIL). Available at: <https://cedilprogramme.org/wp-content/uploads/2019/03/CEDIL-call-spec-for-PoW-2.pdf> [Accessed 4 Sep. 2019].

Chang, H.-J., 2011. Hamlet without the Prince of Denmark: How development has disappeared from today's 'development' discourse. In: S.R. Khan and J. Christiansen, eds. *Towards new developmentalism: market as means rather than master*, Routledge studies in development economics. London ; New York, NY: Routledge.

Chang, W., Díaz-Martin, L., Gopalan, A., Guarnieri, E., Jayachandran, S. and Walsh, C., 2020. What Works to Enhance Women's Agency: Cross-Cutting Lessons From Experimental and Quasi-Experimental Studies. *J-PAL Working Paper*.

Chaudhury, N., Friedman, J. and Onishi, J., 2013. Philippines conditional cash transfer program impact evaluation 2012. *Manila: World Bank Report*, (75533-PH).

Cohen, J., 1992. A power primer. *Psychological Bulletin*, 112(1), pp.155–159. https://doi.org/10.1037/0033-2909.112.1.155.

Conn, K.M., 2017. Identifying effective education interventions in sub-Saharan Africa: A meta-analysis of impact evaluations. *Review of Educational Research*, 87(5), pp.863–898.

Cook, T.D., Campbell, D.T. and Day, A., 1979. *Quasi-experimentation: Design & analysis issues for field settings*. Houghton Mifflin Boston.

Court, J. and Young, J., 2006. Bridging research and policy in international development: context, evidence and links. In: S. Maxwell and D. Stone, eds. *Global knowledge networks and international development: bridges across boundaries*. London; New York: Routledge.pp.18–36.

Cutt, J. and Murray, V., 2000. *Accountability and effectiveness evaluation in nonprofit organizations*. Routledge.

Dalkin, S.M., Greenhalgh, J., Jones, D., Cunningham, B. and Lhussier, M., 2015. What's in a mechanism? Development of a key concept in realist evaluation. *Implementation Science*, 10(1), p.49. https://doi.org/10.1186/s13012-015-0237-x.

Das, J., Do, Q.-T. and Özler, B., 2005. Reassessing conditional cash transfer programs. *The World Bank Research Observer*, 20(1), pp.57–80.

Davey, C., Hargreaves, J., Hassan, S., Cartwright, N., Humphreys, M., Masset, E., Prost, A., Gough, D., Oliver, S. and Bonell, C., 2018. Designing evaluations to provide evidence to inform action in new settings. *CEDIL Inception Paper*, (2).

Davidson, E.J., 2000. Ascertaining causality in theory-based evaluation. *New Directions for Evaluation*, 2000(87), pp.17–26. https://doi.org/10.1002/ev.1178.

Davis, B., Handa, S., Hypher, N., Rossi, N.W., Winters, P. and Yablonski, J., 2016. *From evidence to action: the story of cash transfers and impact evaluation in sub Saharan Africa*. Oxford University Press.

Davis, B., Handa, S., Ruiz -Arranz, M., Stampini, M. and Winters, P., 2002. *Conditionality and the impact of program design on household welfare: comparing two diverse cash transfer programs in rural Mexico*. [online] https://doi.org/10.22004/ag.econ.289104.

De Brauw, A. and Gilligan, D., 2011. Using the regression discontinuity design with implicit partitions: the impacts of Comunidades Solidarias Rurales on schooling in El Salvador. *IFPRI-Discussion Papers*, (1116).

Deans, F. and Ademokun, A., 2013. *Investigating capacity to use evidence: time for a more objective view?* Putting research at the heart of development. Oxford, United Kingdon: INASP.

Deaton, A., 2010. Instruments, Randomization, and Learning about Development. *Journal of Economic Literature*, 48(2), pp.424–455.

Deaton, A. and Cartwright, N., 2018a. Reflections on Randomized Control Trials. *Social Science & Medicine*, 210, pp.86–90. https://doi.org/10.1016/j.socscimed.2018.04.046.

Deaton, A. and Cartwright, N., 2018b. Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, 210, pp.2–21. https://doi.org/10.1016/j.socscimed.2017.12.005.

Denyer, D., Tranfield, D. and van Aken, J.E., 2008. Developing Design Propositions through Research Synthesis. *Organization Studies*, 29(3), pp.393–413. https://doi.org/10.1177/0170840607088020.

DFID, 2011. *Cash Transfers Literature Review*. DFID Policy Division.

DIME, 2015. *What is Impact Evaluation?* [online] World Bank Development Impact Evaluation Group (DIME).p.2. Available at: <http://pubdocs.worldbank.org/en/630921467313243167/What-is-IE-5-6-15-edited.pdf> [Accessed 4 Sep. 2019].

Donnen, P., Brasseur, D., Dramaix, M., Vertongen, F., Zihindula, M., Muhamiriza, M. and Hennart, P., 1998. Vitamin A Supplementation but Not Deworming Improves Growth of Malnourished Preschool Children in Eastern Zaire. *The Journal of Nutrition*, 128(8), pp.1320–1327. https://doi.org/10.1093/jn/128.8.1320.

Donovan, K.P., 2018. The rise of the randomistas: on the experimental turn in international aid. *Economy and Society*, 47(1), pp.27–58. https://doi.org/10.1080/03085147.2018.1432153.

Dossa, R.A., Ategbo, E.-A.D., de Koning, F.L., van Raaij, J.M. and Hautvast, J.G., 2001. Impact of iron supplementation and deworming on growth performance in preschool Beninese children. *European Journal of Clinical Nutrition*, 55(4), p.223.

Dubois, P., De Janvry, A. and Sadoulet, E., 2012. Effects on school enrollment and performance of a conditional cash transfer program in Mexico. *Journal of Labor Economics*, 30(3), pp.555–589.

Duflo, E., 2017. Richard T. Ely Lecture: The Economist as Plumber. *American Economic Review*, 107(5), pp.1–26.

Duflo, E. and Kremer, M., 2005. Use of Randomization in the Evaluation of Development Effectiveness. In: G.K. Pitman, O.N. Feinstein and G.K. Ingram, eds. *Evaluating Development Effectiveness*. Transaction Publishers.

Dunlop, C.A., 2012. Epistemic Communities. [online] Routledge. Available at: <https://ore.exeter.ac.uk/repository/handle/10036/4098> [Accessed 3 Sep. 2019].

Ebenezer, R., Gunawardena, K., Kumarendran, B., Pathmeswaran, A., Jukes, M.C.H., Drake, L.J. and de Silva, N., 2013. Cluster-randomised trial of the impact of school-based deworming and iron supplementation on the cognitive abilities of schoolchildren in Sri Lanka's plantation sector. *Tropical medicine & international health: TM & IH*, 18(8), pp.942–951. https://doi.org/10.1111/tmi.12128.

Edmonds, E.V., 2007. Child labor. *Handbook of development economics*, 4, pp.3607–3709.

Elster, J., 2015. *Explaining social behavior: More nuts and bolts for the social sciences*. Cambridge University Press.

Eyben, R., Guijt, I., Roche, C. and Shutt, C. eds., 2015. *The Politics of Evidence and Results in International Development: Playing the game to change the rules?* [online] Practical Action Publishing Ltd. https://doi.org/10.3362/9781780448855.

Farrington, D.P., Gottfredson, D.C., Sherman, L.W. and Welsh, B.C., 2002. The Maryland Scientific Methods Scale. In: L.W. Sherman, ed. *Evidence-based crime prevention*. London ; New York: Routledge.pp.13–21.

FCDO, 2020. *Smart Rules: Better Programme Delivery*. [online] London, UK: Foreign, Commonwealth and Development Office. Available at: <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/914342/Smart-Rules-External-September_2020.pdf> [Accessed 3 Nov. 2020].

Fenno, R.F., 1986. Observation, Context, and Sequence in the Study of Politics. *American Political Science Review*, 80(01), pp.3–15. https://doi.org/10.2307/1957081.

Ferreira, F.H., Filmer, D. and Schady, N., 2009. *Own and sibling effects of conditional cash transfer programs: Theory and evidence from Cambodia*. The World Bank.

Fiss, P.C., Sharapov, D. and Cronqvist, L., 2013. Opposites Attract? Opportunities and Challenges for Integrating Large-N QCA and Econometric Analysis. *Political Research Quarterly*, 66(1), pp.191–198.

Fiszbein, A. and Schady, N.R., 2009. *Conditional Cash Transfers: Reducing Present and Future Poverty*. World Bank Publications.

Frykman, M., Schwarz, U. von T., Athlin, Å.M., Hasson, H. and Mazzocato, P., 2017. The work is never ending: uncovering teamwork sustainability using realistic evaluation. *Journal of Health Organization and Management*. [online] https://doi.org/10.1108/JHOM-01-2016-0020.

Gaarder, M., 2012. Conditional versus unconditional cash: a commentary. *Journal of Development Effectiveness*, 4(1), pp.130–133.

Gabrielli, A.-F., Montresor, A., Chitsulo, L., Engels, D. and Savioli, L., 2011. Preventive chemotherapy in human helminthiasis: theoretical and operational aspects. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 105(12), pp.683–693.

Galasso, E., 2006. With their effort and one opportunity: Alleviating extreme poverty in Chile. *Unpublished manuscript, World Bank, Washington, DC*.

Garcia, S. and Saavedra, J., 2013. *Educational Impacts and Cost-Effectiveness of Conditional Cash Transfer Programs in Developing Countries: A Meta-Analysis*. [SSRN Scholarly Paper] Rochester, NY: Social Science Research Network. Available at: <http://papers.ssrn.com/abstract=2333946> [Accessed 31 Jan. 2014].

García, S. and Saavedra, J.E., 2017. Educational Impacts and Cost-Effectiveness of Conditional Cash Transfer Programs in Developing Countries: A Meta-Analysis. *Review of Educational Research*, 87(5), pp.921–965. https://doi.org/10.3102/0034654317723008.

Gertler, P.J., Martinez, S., Premand, P., Rawlings, L.B. and Vermeersch, C.M., 2016. *Impact evaluation in practice*. World Bank Publications.

Gitter, S.R. and Barham, B.L., 2009. Conditional cash transfers, shocks, and school enrolment in Nicaragua. *The Journal of Development Studies*, 45(10), pp.1747–1767.

Glaser, B.G. and Strauss, A.L., 2017. *Discovery of Grounded Theory: Strategies for Qualitative Research*. Routledge.

Glasgow, R.E. and Linnan, L.A., 2008. Evaluation of theory-based interventions. *Health behavior and health education: Theory, research, and practice*, 4, pp.487–508.

Gluckman, P., 2013. *The role of evidence in policy formation and implementation*. Auckland, New Zealand: Office of the Prime Minister's Science Advisory Committee.

Greene, J.C., Caracelli, V.J. and Graham, W.F., 1989. Toward a Conceptual Framework for Mixed-Method Evaluation Designs. *Educational Evaluation and Policy Analysis*, 11(3), pp.255–274. https://doi.org/10.3102/01623737011003255.

Grosh, M., del Ninno, C., Tesliuc, E. and Ouerghi, A., 2008. The design and implementation of effective safety nets for protection and promotion. *World Bank, Washington, DC*.

Gupta, M.C. and Urrutia, J.J., 1982. Effect of periodic antiascaris and antigiardia treatment on nutritional status of preschool children. *The American journal of clinical nutrition*, 36(1), pp.79–86.

Haas, P.M., 1992. Introduction: Epistemic Communities and International Policy Coordination. *International Organization*, 46(1), pp.1–35.

Hahn, J., Todd, P. and Van der Klaauw, W., 2001. Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design. *Econometrica*, 69(1), pp.201–209. https://doi.org/10.1111/1468-0262.00183.

Hall, A., 1993. Intestinal parasitic worms and the growth of children. *Transactions of The Royal Society of Tropical Medicine and Hygiene*, 87(3), pp.241–242. https://doi.org/10.1016/0035-9203(93)90108-3.

Harré, R., 1970. *The principles of scientific thinking*. Macmillan.

Harré, R., 1985. *The philosophies of science*. 2nd ed ed. Oxford [Oxfordshire] ; New York: Oxford University Press.

Hawkins, A., 2014. The case for experimental design in realist evaluation. *Learning Communities: International Journal of Learning in Social Contexts*, 14, pp.46–59.

Heckmann, J., 1991. *Randomization and social policy evaluation*. National Bureau of Economic Research.

Hemming, K., Haines, T.P., Chilton, P.J., Girling, A.J. and Lilford, R.J., 2015. The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. *BMJ*, 350, p.h391. https://doi.org/10.1136/bmj.h391.

Hotz, V.J., Imbens, G.W. and Mortimer, J.H., 2005. Predicting the efficacy of future training programs using past experiences at other locations. *Journal of Econometrics*, 125(1–2), pp.241–270.

Humphreys, M. and Jacobs, A.M., 2015. Mixing methods: A Bayesian approach. *American Political Science Review*, 109(4), pp.653–673.

Hyder, A.A., Corluka, A., Winch, P.J., El-Shinnawy, A., Ghassany, H., Malekafzali, H., Lim, M.-K., Mfutso-Bengo, J., Segura, E. and Ghaffar, A., 2011. National policy-makers speak out: are researchers giving them what they need? *Health Policy and Planning*, 26(1), pp.73–82. https://doi.org/10.1093/heapol/czq020.

ICAI, 2020. *About Us*. [online] ICAI. Available at: <https://icai.independent.gov.uk/about-us/> [Accessed 3 Nov. 2020].

Imbens, G., 2018. Understanding and misunderstanding randomized controlled trials: A commentary on Deaton and Cartwright. *Social Science & Medicine*, 210, pp.50–52. https://doi.org/10.1016/j.socscimed.2018.04.028.

Imbens, G.W., 2010. Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009). *Journal of Economic literature*, 48(2), pp.399–423.

Imbens, G.W., 2019. Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics. *arXiv:1907.07271 [stat]*. [online] Available at: <http://arxiv.org/abs/1907.07271> [Accessed 5 Sep. 2019].

Ivers, N.M., Halperin, I.J., Barnsley, J., Grimshaw, J.M., Shah, B.R., Tu, K., Upshur, R. and Zwarenstein, M., 2012. Allocation techniques for balance at baseline in cluster randomized trials: a methodological review. *Trials*, 13(1), p.120. https://doi.org/10.1186/1745-6215-13-120.

Iversen, V. and Lanthorn, H., 2016. *External validity clues in development: On what to look for and how*.

Jamal, F., Fletcher, A., Shackleton, N., Elbourne, D., Viner, R. and Bonell, C., 2015. The three stages of building and testing mid-level theories in a realist RCT: a theoretical and methodological case-example. *Trials*, 16(1), p.466. https://doi.org/10.1186/s13063-015-0980-y.

de Janvry, A., Finan, F. and Sadoulet, E., 2010. *Local Electoral Incentives and Decentralized Program Performance*. [Working Paper] National Bureau of Economic Research. https://doi.org/10.3386/w16635.

Jones, N. and Villar, E., 2008. Situating children in international development policy: challenges involved in successful evidence-informed policy influencing. *Evidence & Policy: A Journal of Research, Debate and Practice*, 4(1), pp.31–51. https://doi.org/10.1332/174426408783477891.

Joseph, S.A., Casapía, M., Montresor, A., Rahme, E., Ward, B.J., Marquis, G.S., Pezo, L., Blouin, B., Maheu-Giroux, M. and Gyorkos, T.W., 2015. The Effect of Deworming on Growth in One-Year-Old Children Living in a Soil-Transmitted Helminth-Endemic Area of Peru: A Randomized Controlled Trial. *PLoS neglected tropical diseases*, 9(10), p.e0004020. https://doi.org/10.1371/journal.pntd.0004020.

Kabeer, N., Piza, C. and Taylor, L., 2012. What are the economic impacts of conditional cash transfer programmes? A systematic review of the evidence. *EPPI CENTRE, SOCIAL SCIENCE RESEARCH UNIT, INSTITUTE OF EDUCATION & UNIVERSITY OF LONDON (eds.)*.

Keck, M.E. and Sikkink, K., 1998. *Activists beyond Borders: Advocacy Networks in International Politics*. Ithaca, NY: Cornell University Press.

Keiser, J. and Utzinger, J., 2008. Efficacy of Current Drugs Against Soil-Transmitted Helminth Infections: Systematic Review and Meta-analysis. *JAMA*, 299(16), pp.1937–1948. https://doi.org/10.1001/jama.299.16.1937.

Kelly, A.H. and McGoey, L., 2018. Facts, power and global evidence: a new empire of truth. *Economy and Society*, 47(1), pp.1–26. https://doi.org/10.1080/03085147.2018.1457261.

Kenny, C., 2021. *We Should Be Spending More of Available Aid in Poorer Countries, Not Less*. [CGD Working Paper] Washington D.C.: Center for Global Development.p.13. Available at:

<https://www.cgdev.org/publication/we-should-be-spending-more-available-aid-poorer-countries-not-less> [Accessed 16 Jan. 2021].

Kuhn, T.S., 1962. *The structure of scientific revolutions*. Chicago ; London: The University of Chicago Press.

Kvangraven, I.H., 2020. Nobel Rebels in Disguise — Assessing the Rise and Rule of the Randomistas. *Review of Political Economy,* 32(3), pp.305–341. https://doi.org/10.1080/09538259.2020.1810886.

Langer, L. and Stewart, R., 2014. What have we learned from the application of systematic review methodology in international development? – a thematic overview. *Journal of Development Effectiveness*, 6(3), pp.236–248. https://doi.org/10.1080/19439342.2014.919013.

Leeuw, F.L. and Vaessen, J., 2009. Impact evaluations and development: NONIE guidance on impact evaluation.

Leijten, P., Melendez-Torres, G.J., Knerr, W. and Gardner, F., 2016. Transported Versus Homegrown Parenting Interventions for Reducing Disruptive Child Behavior: A Multilevel Meta-Regression Study. *Journal of the American Academy of Child & Adolescent Psychiatry*, 55(7), pp.610–617. https://doi.org/10.1016/j.jaac.2016.05.003.

Levy, S., 2007. *Progress Against Poverty: Sustaining Mexico's Progresa-Oportunidades Program*. Brookings Institution Press.

Lewin, S., Glenton, C. and Oxman, A.D., 2009. Use of qualitative methods alongside randomised controlled trials of complex healthcare interventions: methodological study. *BMJ*, 339(sep10 1), pp.b3496–b3496. https://doi.org/10.1136/bmj.b3496.

Lewis, D., 2001. *Counterfactuals*. Wiley.

Littell, J.H., Corcoran, J. and Pillai, V., 2008. *Systematic Reviews and Meta-Analysis*. Oxford University Press, USA.

Mackie, J.L., 1965. Causes and Conditions. *American Philosophical Quarterly*, 2(4), pp.245–264.

Maidment, I., Booth, A., Mullan, J., McKeown, J., Bailey, S. and Wong, G., 2017. Developing a framework for a novel multi-disciplinary, multi-agency intervention(s), to improve medication management in community-dwelling older people on complex medication regimens (MEMORABLE)—a realist synthesis. *Systematic Reviews*, 6(1), p.125. https://doi.org/10.1186/s13643-017-0528-1.

Majid, M.F., Kang, S.J. and Hotez, P.J., 2019. Resolving" worm wars": An extended comparison review of findings from key economics and epidemiological studies. *PLoS neglected tropical diseases*, 13(3), p.e0006940.

Mallett, R., Hagen-Zanker, J., Slater, R. and Duvendack, M., 2012. The benefits and challenges of using systematic reviews in international development research. *Journal of Development Effectiveness*, 4(3), pp.445–455. https://doi.org/10.1080/19439342.2012.711342.

Maluccio, J.A., Murphy, A. and Regalia, F., 2010. Does supply matter? Initial schooling conditions and the effectiveness of conditional cash transfers for grade progression in Nicaragua. *Journal of development effectiveness*, 2(1), pp.87–116.

Marchal, B., Westhorp, G., Wong, G., Van Belle, S., Greenhalgh, T., Kegels, G. and Pawson, R., 2013. Realist RCTs of complex interventions – An oxymoron. *Social Science & Medicine*, 94, pp.124–128. https://doi.org/10.1016/j.socscimed.2013.06.025.

Masset, E. and White, H., 2019. *Designing evaluations to provide evidence to inform action in new settings*. CEDIL Working Paper. [online] London, UK: Centre for Excellence in Development Impact and Learning (CEDIL). Available at: <https://cedilprogramme.org/wp-content/uploads/2019/07/Agenda-Paper-on-template.pdf> [Accessed 4 Sep. 2019].

Maxwell, S. and Stone, D. eds., 2006. Global knowledge networks and international development: bridges across boundaries. In: *Global knowledge networks and international development: bridges across boundaries*. London; New York: Routledge.pp.1–17.

Mayoux, L., 2006. Quantitative, qualitative or participatory? Which method, for what and when? In: *Doing Development Research*. Thousand Oaks: Sage Publications.pp.115–129.

Mazzocato, P., Savage, C., Brommels, M., Aronsson, H. and Thor, J., 2010. Lean thinking in healthcare: a realist review of the literature. *Quality & Safety in Health Care*, 19(5), pp.376–382. https://doi.org/10.1136/qshc.2009.037986.

McShane, B.B., Gal, D., Gelman, A., Robert, C. and Tackett, J.L., 2019. Abandon Statistical Significance. *The American Statistician*, 73(sup1), pp.235–245. https://doi.org/10.1080/00031305.2018.1527253.

Merton, R.K. and Merton, R.C., 1968. *Social Theory and Social Structure*. Simon and Schuster.

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

Moore, G.F., Audrey, S., Barker, M., Bond, L., Bonell, C., Hardeman, W., Moore, L., O'Cathain, A., Tinati, T. and Wight, D., 2015. Process evaluation of complex interventions: Medical Research Council guidance. *bmj*, 350, p.h1258.

Morse, J.M., 1995. The Significance of Saturation. *Qualitative Health Research*, 5(2), pp.147–149. https://doi.org/10.1177/104973239500500201.

Nolan, M. and Grant, G., 1992. Mid-range theory building and the nursing theory-practice gap: a respite care case study. *Journal of Advanced Nursing*, 17(2), pp.217–223. https://doi.org/10.1111/j.1365-2648.1992.tb01876.x.

Oakes, J.M., 2018. The tribulations of trials: A commentary on Deaton and Cartwright. *Social Science & Medicine*, 210, pp.57–59. https://doi.org/10.1016/j.socscimed.2018.04.026.

Ogden, T.N., 2016. *Experimental Conversations: Perspectives on Randomized Trials in Development Economics*. MIT Press.

Oya, C., Schaefer, F. and Skalidou, D., 2018. The effectiveness of agricultural certification in developing countries: A systematic review. *World Development*, 112, pp.282–312. https://doi.org/10.1016/j.worlddev.2018.08.001.

Pawson, R., 2000. Middle-range realism. *European Journal of Sociology / Archives Européennes de Sociologie*, 41(2), pp.283–325. https://doi.org/10.1017/S0003975600007050.

Pawson, R., Greenhalgh, T., Harvey, G. and Walshe, K., 2004. *Realist synthesis: an introduction*. Research Methods. ESRC.

Pawson, R. and Tilley, N., 1997. *Realistic evaluation*. London ; Thousand Oaks, Calif: Sage.

Pearl, J. and Bareinboim, E., 2014. External Validity: From Do-Calculus to Transportability Across Populations. *Statistical Science*, 29(4), pp.579–595. https://doi.org/10.1214/14-STS486.

Pearl, J. and Mackenzie, D., 2018. *The book of why: the new science of cause and effect*. Basic Books.

Petticrew, M. and Roberts, H., 2003. Evidence, hierarchies, and typologies: horses for courses. *Journal of Epidemiology & Community Health*, 57(7), pp.527–529. https://doi.org/10.1136/jech.57.7.527.

Picciotto, R., 2012. Experimentalism and development evaluation: Will the bubble burst? *Evaluation*, 18(2), pp.213–229. https://doi.org/10.1177/1356389012440915.

Pitman, E.J.G., 1937. Significance Tests Which May be Applied to Samples From any Populations. *Supplement to the Journal of the Royal Statistical Society*, 4(1), pp.119–130. https://doi.org/10.2307/2984124.

Porter, S. and O'Halloran, P., 2012. The use and limitation of realistic evaluation as a tool for evidence-based practice: a critical realist perspective: The use and limitation of realistic evaluation. *Nursing Inquiry*, 19(1), pp.18–28. https://doi.org/10.1111/j.1440-1800.2011.00551.x.

Pritchett, L. and Sandefur, J., 2013. *Context Matters for Size: Why External Validity Claims and Development Practice Don't Mix*. CGD Working Papers. Washington D.C.: Center for Global Development.

Pritchett, L. and Sandefur, J., 2015. Learning from Experiments When Context Matters. *American Economic Review*, 105(5), pp.471–475. https://doi.org/10.1257/aer.p20151016.

Puttick, R., 2018. *Mapping the Standards of Evidence used in UK social policy*. Alliance for Useful Evidence. Available at: https://media. nesta. org. uk ….

Ragin, C., 1987. *The Comparative Method: Moving Beyond Qualitative and Quantitative Strategies.* University of California Press.

RAMESES, 2014. *Quality Standards for Realist Synthesis (for researchers and peer-reviewers)*. [online] The RAMESES Project. Available at: <http://www.ramesesproject.org/media/RS_qual_standards_researchers.pdf> [Accessed 15 May 2017].

Raudenbush, S.W., Martinez, A. and Spybrook, J., 2007. Strategies for improving precision in group-randomized experiments. *Educational Evaluation and Policy Analysis*, 29(1), pp.5–29.

Ravallion, M., 2009. Evaluation in the Practice of Development. *The World Bank Research Observer,* 24(1), pp.29–53. https://doi.org/10.1093/wbro/lkp002.

Ravallion, M., 2018. Should the Randomistas (Continue to) Rule? *Center for Global Development Working Paper*, 492.

Ravetz, J., 1995. *Scientific Knowledge and Its Social Problems*. Reprint edition ed. New Brunswick, N.J: Transaction Publishers.

Rogers, P.J., 2007. Theory-Based Evaluation : Reflections Ten Years On.

Rubin, H. and Rubin, I., 2005. *Qualitative Interviewing (2nd ed.): The Art of Hearing Data*. [online] 2455 Teller Road, Thousand Oaks California 91320 United States: SAGE Publications, Inc. https://doi.org/10.4135/9781452226651.

Rubio-Codina, M., 2010. Intra-household time allocation in rural Mexico: Evidence from a randomized experiment. In: *Child labor and the transition between school and work*. Emerald Group Publishing Limited.pp.219–257.

Rycroft-Malone, J., McCormack, B., Hutchinson, A.M., DeCorby, K., Bucknall, T.K., Kent, B., Schultz, A., Snelgrove-Clarke, E., Stetler, C.B., Titler, M., Wallin, L. and Wilson, V., 2012. Realist synthesis: illustrating the method for implementation research. *Implementation Science*, 7, p.33. https://doi.org/10.1186/1748-5908-7-33.

Sabet, S.M. and Brown, A.N., 2018. Is impact evaluation still on the rise? The new trends in 2010–2015. *Journal of Development Effectiveness*, 10(3), pp.291–304. https://doi.org/10.1080/19439342.2018.1483414.

Sanderson, I., 2000. Evaluation in Complex Policy Systems. *Evaluation*, 6(4), pp.433–454. https://doi.org/10.1177/13563890022209415.

Sarkar, N.R., Anwar, K.S., Biswas, K.B. and Mannan, M.A., 2002. Effect of deworming on nutritional status of ascaris infested slum children of Dhaka, Bangladesh. *Indian pediatrics*, 39(11), pp.1021–1026.

Savedoff, W., 2014. *3ie Enters a New Phase and Thank You, Howard White*. [online] Center For Global Development. Available at: <https://www.cgdev.org/blog/3ie-enters-new-phase-and-thank-you-howard-white> [Accessed 3 Sep. 2019].

Sayer, R.A., 1992. *Method in social science: a realist approach*. 2nd ed ed. London ; New York: Routledge.

Schady, N. and Araujo, M.C., 2008. Cash Transfers, Conditions, and School Enrollment in Ecuador. *Economía*, 8(2), pp.43–70. https://doi.org/10.1353/eco.0.0004.

Shadish, W.R., Cook, T.D. and Campbell, D.T., 2002. *Experimental and quasi-experimental designs for generalized causal inference*. Belmont, CA: Wadsworth Cengage Learning.

Silva, N. de, Ahmed, B.-N., Casapia, M., Silva, H.J. de, Gyapong, J., Malecela, M. and Pathmeswaran, A., 2015. Cochrane Reviews on Deworming and the Right to a Healthy, Worm-Free Life. *PLOS Neglected Tropical Diseases*, 9(10), p.e0004203. https://doi.org/10.1371/journal.pntd.0004203.

Simmons, J.P., Nelson, L.D. and Simonsohn, U., 2011. False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, 22(11), pp.1359–1366. https://doi.org/10.1177/0956797611417632.

Singal, A.G., Higgins, P.D.R. and Waljee, A.K., 2014. A Primer on Effectiveness and Efficacy Trials. *Clinical and Translational Gastroenterology*, 5(1), p.e45. https://doi.org/10.1038/ctg.2013.13.

Smillie, W.G. and Augustine, D.L., 1926. HOOKWORM INFESTATION: THE EFFECT OF VARYING INTENSITIES ON THE PHYSICAL CONDITION OF SCHOOL CHILDREN. *American Journal of Diseases of Children*, 31(2), pp.151–168. https://doi.org/10.1001/archpedi.1926.04130020003001.

Smillie, W.G. and Spencer, C.R., 1926. Mental retardation in school children infested with hookworms. *Journal of Educational Psychology*, 17(5), pp.314–321. https://doi.org/10.1037/h0072931.

Snilstveit, B., Oliver, S. and Vojtkova, M., 2012. Narrative approaches to systematic review and synthesis of evidence for international development policy and practice. *Journal of Development Effectiveness*, 4(3), pp.409–429. https://doi.org/10.1080/19439342.2012.710641.

Snilstveit, B., Stevenson, J., Phillips, D., Vojtkova, M., Gallagher, E., Schmidt, T., Jobse, H., Geelen, M., Pastorello, M.G. and Eyers, J., 2015. Interventions for improving learning outcomes and access to education in low-and middle-income countries: a systematic review. *The Campbell Collaboration*.

Soanes, C. and Stevenson, A. eds., 2004. *Concise Oxford English dictionary*. 11. ed ed. Oxford: Oxford Univ. Press.

Speich, B., Knopp, S., Mohammed, K.A., Khamis, I.S., Rinaldi, L., Cringoli, G., Rollinson, D. and Utzinger, J., 2010. Comparative cost assessment of the Kato-Katz and FLOTAC techniques for soil-transmitted helminth diagnosis in epidemiological surveys. *Parasites & Vectors*, 3(1), p.71. https://doi.org/10.1186/1756-3305-3-71.

Stame, N., 2004. Theory-Based Evaluation and Types of Complexity. *Evaluation*, 10(1), pp.58–76. https://doi.org/10.1177/1356389004043135.

Stephenson, L.S., Latham, M.C. and Ottesen, E.A., 2000. Malnutrition and parasitic helminth infections. *Parasitology*, 121(S1), pp.S23–S38.

Stern, E., 2015. Impact Evaluation: A guide for commissioners and managers. *BOND, May*.

Stern, E., Stame, N., Mayne, J., Forss, K., Davies, R. and Befani, B., 2012. Broadening the range of designs and methods for impact evaluations. *Department for International Development Working Paper*, 38.

Stevens, S.S., 1946. On the Theory of Scales of Measurement. *Science*, 103(2684), pp.677–680. https://doi.org/10.1126/science.103.2684.677.

Stuart, E.A., 2010. Matching methods for causal inference: A review and a look forward. *Statistical science : a review journal of the Institute of Mathematical Statistics*, 25(1), pp.1–21. https://doi.org/10.1214/09-STS313.

Taylor, S.J. and Bogdan, R., 1998. *Introduction to Qualitative Research Methods: The Search for Meanings*. 3 edition ed. New York: John Wiley & Sons, Inc.

Tong, A., Sainsbury, P. and Craig, J., 2007. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. *International Journal for Quality in Health Care*, 19(6), pp.349–357. https://doi.org/10.1093/intqhc/mzm042.

Tukey, J.W., 1977. *Exploratory Data Analysis*. Addison-Wesley series in behavioral science. quantitative methods. Reading, Mass. ; London: Addison-Wesley.

Victora, C.G., Onis, M. de, Hallal, P.C., Blössner, M. and Shrimpton, R., 2010. Worldwide Timing of Growth Faltering: Revisiting Implications for Interventions. *Pediatrics*, 125(3), pp.e473–e480. https://doi.org/10.1542/peds.2009-1519.

Vivalt, E., 2015. Heterogeneous Treatment Effects in Impact Evaluation. *The American Economic Review*, 105(5), pp.467–470.

Vivalt, E., 2016. *How Much Can We Generalize from Impact Evaluations?* Unpublished. Stanford University.

Vivalt, E., 2019. *How Much Can We Generalize from Impact Evaluations?* Unpublished. Stanford University.

Vivalt, E., 2020. How Much Can We Generalize From Impact Evaluations? *Journal of the European Economic Association*, 18(6), pp.3045–3089. https://doi.org/10.1093/jeea/jvaa019.

Waddington, H., Masset, E. and Jimenez, E., 2018. What have we learned after ten years of systematic reviews in international development? *Journal of Development Effectiveness*, 10(1), pp.1–16. https://doi.org/10.1080/19439342.2018.1441166.

Webber, S. and Prouse, C., 2018. The New Gold Standard: The Rise of Randomized Control Trials and Experimental Development. *Economic Geography*, 94(2), pp.166–187. https://doi.org/10.1080/00130095.2017.1392235.

Weiss, C., 2009. Foreword. In: *Knowledge to policy: making the most of development research*. Los Angeles : Ottawa: SAGE ; International Development Research Centre.pp.ix–xiii.

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

Weiss, R.S., 1995. *Learning From Strangers: the Art and Method of Qualitative Interview Studies*. [online] Riverside: Free Press. Available at: <https://www.overdrive.com/search?q=4BD4BBBE-6A88-4DE6-93DC-B4915F7DA0D8> [Accessed 3 Sep. 2019].

Welch, V.A., Ghogomu, E., Hossain, A., Awasthi, S., Bhutta, Z.A., Cumberbatch, C., Fletcher, R., McGowan, J., Krishnaratne, S. and Kristjansson, E., 2017. Mass deworming to improve developmental health and wellbeing of children in low-income and middle-income countries: a systematic review and network meta-analysis. *The Lancet Global Health*, 5(1), pp.e40–e50.

Wenger, E., 1998. *Communities of Practice: Learning, Meaning, and Identity*. Cambridge University Press.

Westhorp, G., 2014. *Realist impact evaluation: an introduction*. Overseas Development Institute, the Australian Department of Foreign Affairs and Trade, BetterEvaluation.

White, H., 2009. Theory-based impact evaluation: principles and practice. *Journal of Development Effectiveness*, 1(3), pp.271–284. https://doi.org/10.1080/19439340903114628.

White, H., 2010. A Contribution to Current Debates in Impact Evaluation. *Evaluation*, 16(2), pp.153–164. https://doi.org/10.1177/1356389010361562.

White, H., 2011. *An introduction to the use of randomized controlled trials to evaluate development interventions*. International Initiative for Impact Evaluation.

White, H., 2018. Theory-based systematic reviews. *Journal of Development Effectiveness*, 10(1), pp.17–38. https://doi.org/10.1080/19439342.2018.1439078.

Woolcock, M., 2013. Using case studies to explore the external validity of ‘complex’development interventions. *Evaluation*, 19(3), pp.229–248.

Worrall, J., 2007. Evidence in Medicine and Evidence-Based Medicine. *Philosophy Compass*, 2(6), pp.981–1022. https://doi.org/10.1111/j.1747-9991.2007.00106.x.

Wright, K., 1997. *Knowledge and Expertise in European Conventional Arms Control Negotiations: An Epistemic Community?*. University of Essex.

Yeung, H.W.C., 1997. Critical realism and realist research in human geography: A method or a philosophy in search of a method? *Progress in Human Geography*, 21(1), pp.51–74. https://doi.org/10.1191/030913297668207944.

Zhang, L., Luo, R., Medina, C.A., Liu, C., Rozelle, C.S. and Chen, Y., 2017. *Breaking the cycle of infection: an impact evaluation of three strategies to control intestinal parasites and improve human capital in rural China*. 3ie Grantee Final Report. New Delhi: International Initiative for Impact Evaluation (3ie).

Appendices

Appendix A: Evaluations included in the set for Case One

Akresh, R., De Walque, D. and Kazianga, H., 2013. Cash transfers and child schooling: evidence from a randomized evaluation of the role of conditionality. The World Bank.

Amarante, V., Ferrando, M. and Vigorito, A., 2013. Teenage School Attendance and Cash Transfers: An Impact Evaluation of PANES. Economía, 14(1), pp.61–96.

Angelucci, M., De Giorgi, G., Rangel, M.A. and Rasul, I., 2010. Family networks and school enrolment: Evidence from a randomized social experiment. *Journal of Public Economics,* 94(3), pp.197–221. https://doi.org/10.1016/j.jpubeco.2009.12.002.

Armand, A. and Carneiro, P., 2018. *Impact evaluation of the conditional cash transfer program for secondary school attendance in Macedonia*. 3ie Impact Evaluation Report. International Initiative for Impact Evaluation (3ie).p.48.

Arráiz, I. and Rozo, S., 2011. Same bureaucracy, different outcomes in human capital? How indigenous and rural non-indigenous areas in Panama responded to the CCT. *How Indigenous and Rural Non-Indigenous Areas in Panama Responded to the CCT (May 1, 2011). Office of Evaluation and Oversight Working Paper*, (03/11).

Attanasio, O., Fitzsimons, E., Gomez, A., Gutierrez, M.I., Meghir, C. and Mesnard, A., 2010. Children's schooling and work in the presence of a conditional cash transfer program in rural Colombia. *Economic development and cultural change*, 58(2), pp.181–210.

Baird, S., Chirwa, E., McIntosh, C. and Özler, B., 2010. The short-term impacts of a schooling conditional cash transfer program on the sexual behavior of young women. *Health economics*, 19(S1), pp.55–68.

Baird, S.J., Chirwa, E., Hoop, J. de and Özler, B., 2013. *Girl Power: Cash Transfers and Adolescent Welfare. Evidence from a Cluster-Randomized Experiment in Malawi*. [online] National Bureau of Economic Research. https://doi.org/10.3386/w19479.

Barrera-Osorio, F., Bertrand, M., Linden, L.L. and Perez-Calle, F., 2008. *Conditional Cash Transfers in Education Design Features, Peer and Sibling Effects Evidence from a Randomized Experiment in Colombia*. [Working Paper] National Bureau of Economic Research. https://doi.org/10.3386/w13890.

Behrman, J.R., Parker, S.W. and Todd, P.E., 2004. Medium-term effects of the Oportunidades program package, including nutrition, on education of rural children age 0-8 in 1997. *Unpublished manuscript*.

Benedetti, F., Ibarrarán, P. and McEwan, P.J., 2016. Do education and health conditions matter in a large cash transfer? Evidence from a Honduran experiment. *Economic Development and Cultural Change*, 64(4), pp.759–793.

Benhassine, N., Devoto, F., Duflo, E., Dupas, P. and Pouliquen, V., 2015. Turning a Shove into a Nudge? A “Labeled Cash Transfer” for Education. *American Economic Journal: Economic Policy*, 7(3), pp.86–125. https://doi.org/10.1257/pol.20130225.

Chaudhury, N., Friedman, J. and Onishi, J., 2013. Philippines conditional cash transfer program impact evaluation 2012. *Manila: World Bank Report*, (75533-PH).

Davis, B., Handa, S., Ruiz -Arranz, M., Stampini, M. and Winters, P., 2002. *Conditionality and the impact of program design on household welfare: comparing two diverse cash transfer programs in rural Mexico*. [online] https://doi.org/10.22004/ag.econ.289104.

De Brauw, A. and Gilligan, D., 2011. *Using the regression discontinuity design with implicit partitions: The impacts of comunidades solidarias rurales on schooling in El Salvador*. International Food Policy Research Institute (IFPRI).

De Brauw, A., Gilligan, D.O., Hoddinott, J. and Roy, S., 2015. The impact of Bolsa Família on schooling. *World Development*, 70, pp.303–316.

De Janvry, A., Finan, F. and Sadoulet, E., 2006. *Evaluating Brazil’s Bolsa Escola program: Impact on schooling and municipal roles*. [Working Paper] Available at: <https://are.berkeley.edu/~esadoulet/papers/BolsaEscolaReport6-6.pdf> [Accessed 10 Apr. 2021].

Dubois, P., De Janvry, A. and Sadoulet, E., 2012. Effects on school enrollment and performance of a conditional cash transfer program in Mexico. *Journal of Labor Economics*, 30(3), pp.555–589.

Edmonds, E.V. and Schady, N., 2012. Poverty alleviation and child labor. *American Economic Journal: Economic Policy*, 4(4), pp.100–124.

Edmonds, E.V. and Shrestha, M., 2014. You get what you pay for: Schooling incentives and child labor. *Journal of Development Economics*, 111, pp.196–211.

Evans, D., Hausladen, S., Kosec, K. and Reese, N., 2014. *Community-based conditional cash transfers in Tanzania: Results from a randomized trial*. The World Bank.

Ferré, C. and Sharif, I., 2014. *Can conditional cash transfers improve education and nutrition outcomes for poor children in Bangladesh? Evidence from a pilot project*. The World Bank.

Ferreira, F.H., Filmer, D. and Schady, N., 2009. *Own and sibling effects of conditional cash transfer programs: Theory and evidence from Cambodia*. The World Bank.

Ferro, A.R., Kassouf, A.L. and Levison, D., 2010. The impact of conditional cash transfer programs on household work decisions in Brazil. In: *Child labor and the transition between school and work*. Emerald Group Publishing Limited.

Fuwa, N., 2001. *The Net Impact of the Female Secondary School Stipend Program in Bangladesh*. [MPRA Paper] Available at: <https://mpra.ub.uni-muenchen.de/23402/> [Accessed 10 Apr. 2021].

Galasso, E., 2006. With their effort and one opportunity: Alleviating extreme poverty in Chile. *Unpublished manuscript, World Bank, Washington, DC.*

Galiani, S. and McEwan, P.J., 2013. The heterogeneous impact of conditional cash transfers. *Journal of Public Economics*, 103, pp.85–96.

Garcia, S. and Hill, J., 2010. Impact of conditional cash transfers on children's school achievement: evidence from Colombia. *Journal of Development Effectiveness*, 2(1), pp.117–137.

Gitter, S.R. and Barham, B.L., 2009. Conditional cash transfers, shocks, and school enrolment in Nicaragua. *The Journal of Development Studies*, 45(10), pp.1747–1767.

Glewwe, P. and Kassouf, A.L., 2008. The impact of the Bolsa Escola/Família conditional cash transfer program on enrollment, grade promotion and drop out rates in Brazil. *Anais do XXXVIII Encontro Nacional de Economia*.

Ham González, A., 2010. *The effect of conditional cash transfers on educational opportunities: experimental evidence from Latin America*. Documentos de Trabajo del CEDLAS. [Working Paper] Available at: <https://www.researchgate.net/profile/Andres-Ham/publication/254406362_The_Effect_of_Conditional_Cash_Transfers_on_Educational_Opportunities_-_Experimental_Evidence_from_Latin_America/links/54b518d50cf28ebe92e4be80/The-Effect-of-Conditional-Cash-Transfers-on-Educational-Opportunities-Experimental-Evidence-from-Latin-America.pdf> [Accessed 10 Apr. 2021].

Handa, S., Park, M., Darko, R.O., Osei-Akoto, I., Davis, B. and Daidone, S., 2014. *Livelihood empowerment against poverty program impact evaluation*. 3ie Grantee Final Report. New Delhi: International Initiative for Impact Evaluation (3ie).

de Janvry, A., Finan, F. and Sadoulet, E., 2010. *Local Electoral Incentives and Decentralized Program Performance*. [Working Paper] National Bureau of Economic Research. https://doi.org/10.3386/w16635.

Li, F., Song, Y., Yi, H., Wei, J., Zhang, L., Shi, Y., Chu, J., Johnson, N., Loyalka, P. and Rozelle, S., 2017. The impact of conditional cash transfers on the matriculation of junior high school students into rural China's high schools. *Journal of Development Effectiveness*, 9(1), pp.41–60.

Maluccio, J.A., Murphy, A. and Regalia, F., 2010. Does supply matter? Initial schooling conditions and the effectiveness of conditional cash transfers for grade progression in Nicaragua. *Journal of development effectiveness*, 2(1), pp.87–116.

Mo, D., Zhang, L., Yi, H., Luo, R., Rozelle, S. and Brinton, C., 2013. School dropouts and conditional cash transfers: evidence from a randomised controlled trial in rural China's junior high schools. *The Journal of Development Studies*, 49(2), pp.190–207.

O'Brien, C., Marzi, M., Pellerano, L. and Visram, A., 2016. *The Impact of BOTA's Conditional Cash Transfer (CCT) Programme*. KAZAKHSTAN: EXTERNAL EVALUATION OF BOTA PROGRAMMES. Oxford, United Kingdon: Oxford Policy Management.

Olinto, P. and Souza, P.Z. de, 2005. *An impact evaluation of the conditional cash transfers to education in praf: an experimental approach*. [Master's Dissertation] Available at: <http://bibliotecadigital.fgv.br/dspace/handle/10438/79> [Accessed 10 Apr. 2021].

Perova, E., 2010. *Three essays on intended and not intended impacts of conditional cash transfers*. PhD Thesis. UC Berkeley.

Rubio-Codina, M., 2010. Intra-household time allocation in rural Mexico: Evidence from a randomized experiment. In: *Child labor and the transition between school and work*. Emerald Group Publishing Limited.pp.219–257.

Ward, P., Hurrell, A., Visram, A., Riemenschneider, N., Pellerano, L., O’Brien, C., MacAuslan, I. and Willis, J., 2010. *Cash Transfer Programme for Orphans and Vulnerable Children (CT-OVC), Kenya*. [Operational and Impact Evaluation - Final report] Oxford, United Kingdon: Oxford Policy Management. Available at: <https://d1wqtxts1xzle7.cloudfront.net/53073310/Kenya_2010-019_OPM_CT-OVC_evaluation_report_july2010-final.pdf?1494444482=&response-content-disposition=inline%3B+filename%3DCASH_TRANSFER_PROGRAMME_FOR_ORPHANS_AND.pdf&Expires=1618076185&Signature=Uj8aFnL6~Pnmmgd5vjZ1JV96EV9Hi~NQhsVVlR2SVWZpQMFYLR3iAUPgGtw~lI7fOJwDiufufnouWxHVeJQ~jPUf~5qRmISKlIapEC4JM6-qxqY3KefCSL5AzO5tFWr2m8EikZdCPqXEQWuEjV7wyT0e6aum4GE2fkQZajeTNYVrc4SEY~BiGxyMFsPxVMOX31aqia9SZ8K~Q46yQ3eBo86xy-Wr8JEwLhW1XN69y0kyIJBzlLlgjsXYqdjcGzXJAaWCQO1q2n72sTrdGMaAl5g~1bkHVNbY7HGkAOkK53YwAAY9Q4gPcy2pUutedMcvom-CkgtirlPJnKyq2tDd~w__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA> [Accessed 10 Apr. 2021].

Appendix B: Evaluations included in the set for Case Two

Alderman, H., Konde-Lule, J., Sebuliba, I., Bundy, D. and Hall, A., 2006. Effect on weight gain of routinely giving albendazole to preschool children during child health days in Uganda: cluster randomised controlled trial. *bmj*, 333(7559), p.122.

Awasthi, S. and Pande, V.K., 2001. Six-monthly de-worming in infants to study effects on growth. *The Indian Journal of Pediatrics*, 68(9), pp.823–827.

Awasthi, S., Peto, R., Pande, V.K., Fletcher, R.H., Read, S. and Bundy, D.A.P., 2008. Effects of Deworming on Malnourished Preschool Children in India: An Open-Labelled, Cluster-Randomized Trial. *PLOS Neglected Tropical Diseases*, 2(4), p.e223. https://doi.org/10.1371/journal.pntd.0000223.

Awasthi, S., Peto, R., Read, S., Richards, S.M., Pande, V., Bundy, D. and DEVTA (Deworming and Enhanced Vitamin A) team, 2013. Population deworming every 6 months with albendazole in 1 million pre-school children in North India: DEVTA, a cluster-randomised trial. *Lancet (London, England)*, 381(9876), pp.1478–1486. https://doi.org/10.1016/S0140-6736(12)62126-6.

Baird, S., Hicks, J.H., Kremer, M. and Miguel, E., 2015. *Worms at Work: Long-run Impacts of a Child Health Investment*. [Working Paper] National Bureau of Economic Research. https://doi.org/10.3386/w21428.

Bobonis, G.J., Miguel, E. and Puri-Sharma, C., 2006. Anemia and school participation. *Journal of Human resources*, 41(4), pp.692–721.

Donnen, P., Brasseur, D., Dramaix, M., Vertongen, F., Zihindula, M., Muhamiriza, M. and Hennart, P., 1998. Vitamin A Supplementation but Not Deworming Improves Growth of

Malnourished Preschool Children in Eastern Zaire. *The Journal of Nutrition*, 128(8), pp.1320–1327. https://doi.org/10.1093/jn/128.8.1320.

Dossa, R.A., Ategbo, E.-A.D., de Koning, F.L., van Raaij, J.M. and Hautvast, J.G., 2001. Impact of iron supplementation and deworming on growth performance in preschool Beninese children. *European Journal of Clinical Nutrition*, 55(4), p.223.

Ebenezer, R., Gunawardena, K., Kumarendran, B., Pathmeswaran, A., Jukes, M.C.H., Drake, L.J. and de Silva, N., 2013. Cluster-randomised trial of the impact of school-based deworming and iron supplementation on the cognitive abilities of schoolchildren in Sri Lanka's plantation sector. *Tropical medicine & international health: TM & IH*, 18(8), pp.942–951. https://doi.org/10.1111/tmi.12128.

Garg, R., Lee, L.A., Beach, M.J., Wamae, C.N., Ramakrishnan, U. and Deming, M.S., 2002. Evaluation of the integrated management of childhood illness guidelines for treatment of intestinal helminth infections among sick children aged 2–4 years in western Kenya. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 96(5), pp.543–548.

Gupta, M.C. and Urrutia, J.J., 1982. Effect of periodic antiascaris and antigiardia treatment on nutritional status of preschool children. *The American journal of clinical nutrition*, 36(1), pp.79–86.

Hadju, V., Stephenson, L.S., Mohammed, H.O., Bowman, D.D. and Parker, R.S., 1998. Improvements of growth. Appetite, and physical activity in helminth-infected schoolboys 6 months after single dose of albendazole. *Asia Pacific journal of clinical nutrition*, 7, pp.170–176.

Joseph, S.A., Casapía, M., Montresor, A., Rahme, E., Ward, B.J., Marquis, G.S., Pezo, L., Blouin, B., Maheu-Giroux, M. and Gyorkos, T.W., 2015. The Effect of Deworming on Growth in One-Year-Old Children Living in a Soil-Transmitted Helminth-Endemic Area of Peru: A

Randomized Controlled Trial. *PLoS neglected tropical diseases*, 9(10), p.e0004020. https://doi.org/10.1371/journal.pntd.0004020.

Kruger, M., Badenhorst, C.J., Mansvelt, E.P., Laubscher, J.A. and Benadé, A.S., 1996. Effects of iron fortification in a school feeding scheme and anthelmintic therapy on the iron status and growth of six-to eight-year-old schoolchildren. *Food and Nutrition Bulletin*, 17(1), pp.1–11.

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

Palupi, L., Schultink, W., Achadi, E. and Gross, R., 1997. Effective community intervention to improve hemoglobin status in preschoolers receiving once-weekly iron supplementation. *The American journal of clinical nutrition*, 65(4), pp.1057–1061.

Sarkar, N.R., Anwar, K.S., Biswas, K.B. and Mannan, M.A., 2002. Effect of deworming on nutritional status of ascaris infested slum children of Dhaka, Bangladesh. *Indian pediatrics*, 39(11), pp.1021–1026.

Shally, A., Pande, V.K. and Fletcher, R.H., 2000. Effectiveness and cost-effectiveness of albendazole in improving nutritional status of pre-school children in urban slums. *Indian Pediatrics*, 37(1), pp.19–29.

Stephenson, L.S., Latham, M.C., Kurz, K.M., Kinoti, S.N. and Brigham, H., 1989. Treatment with a single dose of albendazole improves growth of Kenyan schoolchildren with hookworm, Trichuris trichiura, and Ascaris lumbricoides infections. *The American journal of tropical medicine and hygiene*, 41(1), pp.78–87.

Stoltzfus, R.J., Albonico, M., Tielsch, J.M., Chwaya, H.M. and Savioli, L., 1997. School-based deworming program yields small improvement in growth of Zanzibari school children after one year. *The Journal of nutrition*, 127(11), pp.2187–2193.

Thein-Hlaing, Thane-Toe, Than-Saw, Myat-Lay-Kyin and Myint-Lwin, 1991. A controlled chemotherapeutic intervention trial on the relationship between Ascaris lumbricoides infection and malnutrition in children. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 85(4), pp.523–528.

Watkins, W.E. and Pollitt, E., 1996. Effect of removing Ascaris on the growth of Guatemalan schoolchildren. *Pediatrics*, 97(6), pp.871–876.

Zhang, L., Luo, R., Medina, C.A., Liu, C., Rozelle, C.S. and Chen, Y., 2017. *Breaking the cycle of infection: an impact evaluation of three strategies to control intestinal parasites and improve human capital in rural China*. 3ie Grantee Final Report. New Delhi: International Initiative for Impact Evaluation (3ie).

Appendix C: Basic information for research participants

Basic information for research participants

Who I am

Formerly of the Center for Global Development, I am now a second-year PhD student at SOAS, University of London in the department of development studies. My research interests are centred in international development and the philosophy of social science. You can find out a bit more about me and my previous research at mattjuden.com, if you would like.

The purpose of my research

In this part of my PhD research, I'm assessing the views of experts in the evaluation of development interventions at the micro level regarding what counts as 'high quality evidence' of such impacts.

I am attempting to get at the theoretical underpinnings of these views to describe the account of evidence quality at work. Further, for experts employing similar accounts, I'm trying to explore what features of these accounts they're happy with. I'm also interested in whether they perceive weaknesses in these accounts or knowledge puzzles relating to those accounts that need to be worked upon.

This work is intended to allow me to shape the findings of the other strand of my PhD research in order to make it as useful as possible for the relevant experts. This other strand explores how different impact evaluation methods differently explore and report on the contextual factors of relevance to intervention causation. I do this by assessing impact evaluations of the same intervention-outcome pairing but which employ a variety of methods. I then compare both within sets of studies of a given intervention-outcome pairing and between sets of studies of different intervention-outcome pairings. This comparison is used to investigate whether there are any systematic differences between evaluation methods regarding researchers' exploration and reporting of contextual factors of relevance to intervention causation. This empirical work is used to inform the construction of a theoretical framework for thinking about how to facilitate the transportability between contexts of impact evaluation results.

Topics we will discuss

- What makes a high-quality impact evaluation?
- Transportability of results from one context to another – is this important? In what way?
- How can we assess the total available evidence on a given policy question?
- What room for improvement is there in the way that you and your colleagues judge the quality of impact evaluations?
- What's next in impact evaluation?

How your answers will be used

I will record our interview, having asked you to consult and sign a consent form giving more detail about the storage and use of your answers. Your answers will then be transcribed and combined with other answers to create a dataset of opinions of relevant experts. Any reference to, or excerpts or quotations from specific answers reproduced in my PhD thesis or other work will always be anonymised and will contain no identifying data.

Appendix D: Participant consent form

PARTICIPANT CONSENT FORM

Introduction

The purpose of this form is to provide you with information so you can decide whether to participate in this study. Any questions you may have will be answered by the researcher or by the other contact persons provided below. Once you are familiar with the information on the form and have asked any questions you may have, you can decide whether or not to participate. If you agree, please either sign this form or else provide verbal consent

Research title: Towards a multi-dimensional model of evidence quality: assessing development research methods with respect to context

Type of Project PhD Research

Project funder: The Economics and Social Sciences Research Council

Research coordinator: Matthew Juden

m_juden@soas.ac.uk / matt@mattjuden.com

+44 7460 524 377

Purpose of Research: In this part of my PhD research, I'm assessing the views of experts in the evaluation of development interventions at the micro level regarding what counts as 'high quality evidence' of such impacts.

The overall goal of my research is to inform an answer to the research question: On questions of the effectiveness of development policy interventions, can we give a useful, systematic account of the relative merits of evidence generated using different methods that goes beyond internal validity to also consider generalisability?

Reasons for data collection:

You have been selected as an anglophone expert on the evaluation of development interventions at the micro level. I am attempting to sample a broad range of such experts, up to the maximum number interviews that it is practical for me to conduct in the time available to me, or up to a point of conceptual saturation.

Nature of Participation

One 45 minute interview, audio recorded for the purposes of abbreviated transcription of key points.

Risks and Benefits of participation

This strand of my research intends to inform the conception, presentation and dissemination of results from the other strand of my research in order to maximise the relevance and utility of that research to development impact evaluation experts like yourself. Participation should therefore increase the chance that my research as a whole is of interest and relevance to you.

The anonymization of your responses and removal of identifying characteristics prior to publication is intended to minimize any reputational risk for you or your oganisation. Your consent will be sought be separately and specifically for any attributed quote or insight.

Data Sharing:

Participant data will not be shared with any third party except in abbreviated, anonymised form with identifying information removed. Such data sharing is not currently envisaged, though it could be engaged in for the purposes of promoting research integrity.

Countries to which the data may be a transferred:

Data about you gathered in the course of your participation in this project may be transferred to countries or territories outside the European Economic Area for purposes connected with this project and similar future projects, subject to appropriate safeguards to protect the security and confidentiality of your data.

Security measures: Recordings of interviews, transcripts, and the final database of responses will be stored in a secure online repository using strong, two-factor authentication and accessible only to the primary researcher.

Methods of anonymisation: Any reference to, or excerpts or quotations from specific answers reproduced in my PhD thesis or other work will always be anonymised and will contain no identifying data. The only exception will be in the case of attributed quotations from interviews that will be specifically cleared with participants prior to publication. Signing this consent form does not constitute consent for such use of answers given in this interview. Rather, you will be able to give or withhold consent for such use at a later date.

Methods of publication: Your responses will inform research outputs for publication as a PhD thesis, and possibly academic articles and books. Open access to such publications will be sought but cannot be guaranteed due to funding constraints.

Withdrawal of Consent

Please note your participation is voluntary and you may decide to leave the study at any time. You may also refuse to answer specific questions you are uncomfortable with. You may withdraw permission for your data to be used, at any time up to [Researcher to enter date or point in project when it is no longer possible to withdraw consent for use of personal data e.g. when data has been anonymized] in which case notes, transcriptions and recordings will be destroyed. Withdrawal or refusal to participate will not affect your relationship with [Insert name of organization to which research participant belongs if you are doing research in an organization. Remove this statement if not appropriate].

Data Protection Statement

Information about you which is gathered in the course of this research project, once held in the United Kingdom, will be protected by the UK Data Protection Act and will be subject to SOAS's Data Protection Policy. You have the right to request access under the Data Protection Act to the information which SOAS holds about you. Further information about your rights under the Act and how SOAS handles personal data is available on the Data Protection pages of the SOAS website (http://www.soas.ac.uk/infocomp/dpa/index.html), and by contacting the Information Compliance Manager at the following address: Information Compliance Manager, SOAS, Thornhaugh Street, Russell Square, London WC1H 0XG, United Kingdom (e-mail to: dataprotection@soas.ac.uk).

Copyright Statement

By completing this form, you permit SOAS and the research to edit, copy, disseminate, publish (by whatever means) and archive your contribution to this research project in the manner and for the purposes described above. You waive any copyright and other intellectual property rights in your contribution to the project, and grant SOAS and researchers who are involved, a non-exclusive, free, irrevocable, worldwide license to use your contribution for the purposes of this project. If you wish to receive a copy final published research outputs once completed I will happy to provide you with an electronic copy

Contact Information

Telephone No: +44 7460 524 377

Email Address: m_juden@soas.ac.uk / matt@mattjuden.com

Postal Address: SOAS University of London
Thornhaugh Street
Russell Square
London WC1H 0XG
United Kingdom

Alternative contact: Peter Mollinga – pm35@soas.ac.uk

Research Participant Declaration

I confirm that I have read the above information relating to the research project. I freely consent to my information being used in the manner and for the purposes described, and I waive my copyright and other intellectual property rights as indicated. I understand that I may withdraw my consent to participate in the project, and that I should contact the project coordinator if I wish to do so.

Participant Name:

Signature: **Date:**

Researcher Name: Matthew Juden

Signature: **Date:**

PLEASE KEEP THIS FORM FOR FUTURE REFERENCE

Appendix E: Interview guide

Interview guide

Ice-breaking and background info generating

Introduce what I'm doing.

Just for the tape/to get us started…

…what is your name and…

…can you describe your role (at org.)?

Unpack in the direction of assessment of impact evidence

You mentioned…

…does that involve impact evaluations (IEs)?

…does that mean commissioning/synthesising/using IEs?

Guide into specifics of IE quality

What makes a high-quality impact evaluation?

Can you talk more about…? / What do you mean by…? **Until topic exhausted.**

Specifically investigate transferability of results (if this hasn't come up)

Is transferability of results important? How so?

Can you talk more about…? / What do you mean by…? Until topic exhausted.

Attempt to elicit the account of transferability adopted in this research project

You've talked about 'taking transferability' seriously etc. What does this mean for reporting? / What things to evaluations need to report to facilitate transferability?

Come at the same topics from the angle of evidence amalgamation/synthesis

Have you ever assessed the available evidence on a given policy question concerning development at the micro level? What did you do?

What do you think is the best way of **assessing the available evidence**? What should one do?

Unpack any problems

What room for improvement is there in the way that your and colleagues judge the quality of impact evaluations?

You mentioned… …is that a problem? / … can you talk about that more?

Unpack plus points / look to future

You mentioned… etc.

Looking forward, what do you think we'll see more of in impact evaluation / evidence synthesis?

Is there anything else you want to talk about?

Who else should I talk to? (dissimilar opinions/institution)