Realist randomised controlled trials of development interventions in practice: concrete design suggestions to address the problem of external validity

---

Matthew Tom Juden

This dissertation is submitted in partial fulfilment of the requirements for the degree of MSc in Research Methods for International Development of the School of Oriental and African Studies (University of London).

Declaration:	"I undertake that all material presented for examination is my own work and has not been written for me, in whole or in part, by any other persons(s). I also undertake that any quotation or paraphrase from the published or unpublished work of another person has been duly acknowledged in the work which I present for examination."
Signed:	Matthew Tom Juden
World count:	12,850 words
Submission date:	15/09/2014
Acknowledgements	I would like to thank Professor Peter Mollinga and Dr Carlos Oya for helpful conversations during the research and question-framing phase of this dissertation.

CONTENTS

ABSTRACT

Randomised controlled trials (RCTs) of development interventions at the micro level are increasingly common. However, it has been suggested that they suffer from a problem of external validity: their conclusions are not rigorously generalizable. This dissertation examines the strengths and weaknesses of the RCT method and concludes that the problem of external validity is a problem of successionist accounts of causation, not the RCT method. It is argued that the strengths of RCTs can be salvaged through their incorporation into a realist research strategy. The concrete implications for trial design are examined and suggestions made. Barriers to the widespread adoption of this methodology are examined and it is argued that the critiquing of 'exemplar trials' in the mainstream from a realist perspective is a good strategy for overcoming inertia. Two case study trials are critiqued, demonstrating the advantages of realist RCT design incorporating a generative causal model.

1 – INTRODUCTION

> *'Britain has given the world Shakespeare, Newtonian physics, the theory of evolution, parliamentary democracy – and the randomized controlled trial.'*
>
> The British Medical Journal (2001 p.1438) quoted in Worral (2007)

This hyperbolic sentence in the BMJ is typical of the recent enthusiasm surrounding randomised controlled trials (RCTs) of development interventions, especially in the 'complex public health interventions' and 'cash transfers' literatures (Adato *et al.*, 2010; Banerjee and Duflo, 2008; Craig *et al.*, 2008). RCTs are increasingly seen as the 'gold standard' for evidence generation about development interventions at the micro level (Deaton, 2010). However, there is a strong counter-current in the literature: 'realist' social scientists have criticised RCTs for not producing generalizable findings, and philosophers of science appraising 'evidence-based policy' have argued that there is no 'gold standard' method for evidence generation (Cartwright, 2007; Pawson and Tilley, 1997). Recently it has been suggested that the RCT method and realism are reconcilable by designing RCTs so as to contribute to a research strategy based on realist epistemic and ontological assumptions (Bonell *et al.* 2012). However, it is unclear precisely what modifications are required at the design level. This dissertation makes design suggestions based on an iterative engagement with the theoretical literature. It also explores the possibilities for these changes to be implemented. Two mainstream trials are critiqued from a realist perspective in order to motivate researchers to consider the advantages of realist RCTs.

Chapter two explores the strengths and weaknesses of the RCT method and argues that RCTs as typically designed are crippled by the problem of external validity. It is argued that this is more properly thought of as a problem for successionist accounts of causation and that realist RCTs based on a generative account of causation are the solution to this problem. It is further argued that incorporating RCTs into a realist epistemic strategy would undermine 'gold standard' thinking and improve the profile of non-RCT methods. A move beyond 'realism' to 'critical realism' is also argued for. Chapter three explores barriers to the widespread adoption of realist RCTs and argues that the

best strategy to achieve change is to engage with ‘exemplar trials’ at the design level. To that end, concrete suggestions for realist RCT design are made. Chapter four critiques two high-profile trials, having defended the choice of case studies. Chapter five concludes.

2 – A THEORETICAL ARGUMENT FOR REALIST RANDOMISED CONTROLLED TRIALS

2.1 – The appeal of the randomised controlled trial

The reasons behind the recent popularity of RCTs of development interventions are many and varied including complex processes such as the increasing importance of the 'professional evaluation community' (Sanderson, 2000 p.436) and the 'economics imperialism' outlined by Fine (2002). However, it would be implausible to suggest that this popularity has nothing to do with the strengths of the method. This section examines what is meant by a 'randomised controlled trial' and what makes this type of research design powerful and therefore powerfully attractive.

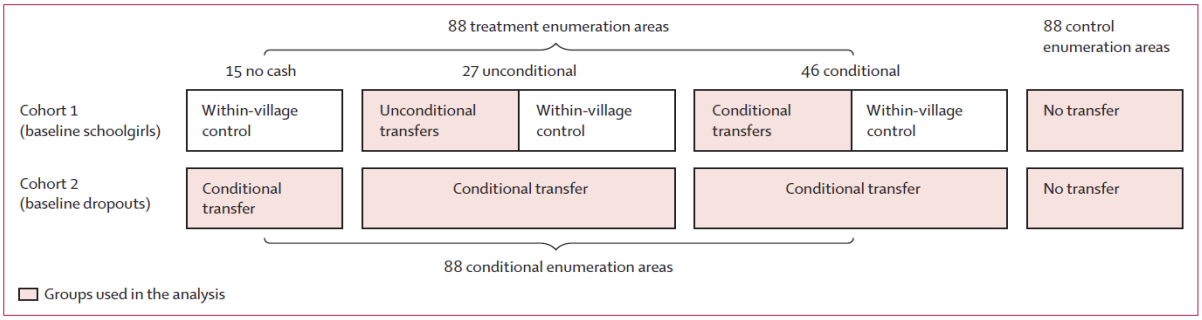
A randomised controlled trial (RCT) is a trial in which subjects are *randomly* assigned to different arms of the trial, one of which is a control group that does not receive the intervention. In the simplest form of RCT, there are two arms: one treatment group and one control group. 'Baseline' measurements of some characteristic(s) of all the subjects are measured before the intervention, and 'endline' measurements of the same characteristic(s) of all the subjects are measured after the intervention has ended. In more complex trials, referred to as 'multiple treatment experiments' by Banerjee and Duflo (2008 p.6), multiple treatment groups are formed and each group receives a different form of the intervention. Banerjee and Duflo (ibid p.4) argue that the key strength of the RCT method is the ability to 'vary one factor at a time and therefore provide "internally" valid estimates of the causal effect' even in the face of 'complex and multiple channels of causality'. This is slightly too generous as RCTs do not allow social scientists to vary any 'factor'. One cannot decide to conduct an experiment in Malawi and vary the 'form of government factor' in such a way as to assign some participants to a multi-party democracy and others to a theocratic regime, for example. Rather, RCTs allow social scientists to vary aspects of the intervention so as to answer counterfactual questions of the form 'what would have happened if the intervention had been different in such and such a way?'

Randomization can occur at the individual level or at a group level, leading to cluster-randomization. The advantage of combining cluster randomization at a group level with randomization at the

individual level is that it allows researchers to measure 'spillover' effects like the difference in mean outcome between control subjects in control clusters and control subjects in treatment clusters. Random assignment can also be combined with non-random assignment: individuals within the sampling population can be non-randomly separated into different cohorts based on some differentiating characteristic of those individuals and then subsequently randomly assigned to different arms of the trial (Baird *et al.*, 2012).

For example, Baird *et al.* (2012) is a cluster-randomised multiple treatment RCT which was conducted in Zomba district, southern Malawi to investigate 'the efficacy of a cash transfer programme to reduce the risk of sexually transmitted infections in young women' (p.1). Their selection began with 176 enumeration areas which were cluster-randomised into 88 treatment areas and 88 control areas. In order to test the effect of conditionality on outcomes, treatment areas were further cluster-randomised to receive conditional or unconditional cash transfers. The study was concerned with two different cohorts: young women enrolled in school at the beginning of the study (as measured by a baseline survey), and those not enrolled at baseline. It was also concerned with two different treatments: an unconditional cash transfer and a conditional cash transfer. Furthermore the study wished to examine spillover effects. In order to permit fine-grained analysis of spillover effects the percentage of subjects enrolled in school at baseline who were assigned to treatment was set at 0%, 33%, 66% or 100% for different subsets of those enumeration areas which had been assigned to treatment. This meant that for 15 treatment enumeration areas only baseline dropouts received the intervention; all baseline schoolgirls were assigned to control. The resultant allocation of subjects to different arms of the trial is shown below in Baird *et al.*'s (ibid) figure 1, reproduced here as figure 1:

Figure 1



(Baird *et al.* 2012, p.1322: Figure 1)

Banerjee and Duflo (2008, pp.9-10) suggest that 'the most important element of the experimental approach' may be the convenience of creating different treatment arms in an RCT as illustrated in figure 1. However, the *unique* strength of the RCT does not lie in facilitating this; it is possible to find or generate observational data in which groups of subjects have received different forms of an intervention or not received any intervention at all. Rather, the unique strength of the RCT lies in the random allocation of subjects to different arms of the trial. Randomised allocation to trial arms is powerful because it allows researchers to isolate the treatment effect from other possible causes of the changes observed in the observed characteristic(s) of subjects in the trial. In a study in which subjects have been non-randomly assigned to treatment or control groups, for example by virtue of living in a particular area, one cannot rule out the possibility that changes observed in subjects' characteristic(s) of interest are due to differences between the treatment and control groups introduced by this selection bias rather than the intervention itself. If subjects are given the choice of whether to participate in the programme or not, then those who choose the intervention may be systematically different from those who do not in ways other than the mere fact of their choice. For example, we can imagine that subjects who choose to take part in a training programme are more motivated and energetic than those who do not and may have had better outcomes even without training. This is a specific example of the problem 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). 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 eliminates selection bias and endogeneity. To put it another way, randomised controlled trials allow researchers to 'examine the counterfactual'; they allow researchers to answer the question 'what would have happened if the intervention had been different in a certain way or had not taken place?' by making this question logically equivalent to the question 'what *did* happen in the relevant arm of the trial?'

Cartwright (2007, pp.13-15) clearly explains that RCTs not only allow researchers to *generate* 'minimally biased estimates of intervention effects', they allow researchers to *deduce* these causal effects using only a small set of assumptions. Paraphrasing to avoid the technical philosophical jargon employed by Cartwright which it is beyond the scope of this dissertation to elaborate upon, these assumptions are:

1) The adoption of some theory of causality that allows the researcher to move from probabilities to causation such as Patrick Suppes' (1970) probabilistic theory of causality or Granger (1969) causality.
    - This implies that 'if the probability of an "outcome" O is greater with a putative cause T than without T once all "confounders" are controlled for in some particular way, that is sufficient for the claim "T causes O" in that particular setting of confounding factors.' (ibid p.12) 'Confounding factors' are causal processes other than T which influence O.
2) That the RCT is an ideal RCT involving 'careful use of statistics to move from frequencies to probabilities, "random" assignment to treatment and control groups, quadruple blinding, careful attention to drop-outs and noncompliance, and so on', meaning that there is no systematic difference between individuals in different arms of the study. (ibid p.15)
    - This means that all 'confounding factors' can be assumed to be equally distributed between different arms of the trial and the difference between mean outcomes in

different arms of the trial can be causally attributed to the intervention, the putative cause T in point 1).

When these assumptions are met, they mean that the RCT constitutes a deductive proof of the causal effect of the intervention in the test population. Cartwright (2007, p.6) calls this kind of argument a 'clincher'. By contrast, the conclusions of an observational study are always open to a fresh challenge that some causally relevant difference between the groups compared has been left out of the analysis. This is what Cartwright (ibid) means by saying that such studies merely 'vouch for' their conclusions rather than clinching them. A different way of capturing this idea is to say that a positive result in an observational study constitutes a necessary but not sufficient condition for establishing the truth of 'T causes O in the test population φ', whereas a positive result in an RCT would be sufficient.

Cartwright's attempt to formalise the appeal of RCTs is an accurate portrayal of the appeal of RCTs for those who implicitly or explicitly adopt a 'successionist' account of causation. The image of causality in assumption 1) is clearly 'successionist' in the sense intended by Harré (1985, p.116) who made the distinction between 'the two great metaphysical theories of causality' by saying:

> *'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 is motivated and underpinned by positivist ontology that is anti-realist about causal powers or 'mechanisms'; it does not posit the existence of causal powers in the world. This anti-realism is based on the apparently sensible observation that we cannot observe causal mechanisms. For example, in *An Enquiry Concerning Human Understanding* Hume (1748, §IV, para.29) states that '[o]ur senses inform us of the colour, weight, and consistence of bread; but neither sense nor reason [for Hume, *a priori* deductive reasoning] can ever inform us of those qualities which

fit it for the nourishment and support of a human body'. Hume goes on to conclude that talking of bread *causing* nourishment is merely a convenient shorthand for the observed regularity or 'constant conjunction' of eating bread and being nourished. Causation then becomes a psychological phenomenon rather than an inherent property of things in the world (Harré, 1985, p.117). Put differently, causation under the successionist account is a relationship between discrete objects which is reducible to their covariance, with the cause temporally preceding the effect. One could attack the RCT's claim to be a 'clinching' deductive method of proving causality by attacking the positivist ontology which motivates successionist thinking about causation, thus undermining assumption 1) above. Harré (1972, 1985) and others (Bhaskar 1975; Sayer 1992) have provided ample ammunition for this attempt. However, attacking successionist theories of causality on ontological grounds is not necessary for the purposes of this dissertation. A successionist picture of causality will for now be granted as reasonable, though in the next section it will be shown that such an account is unable to provide any satisfactory way out of the problem of external validity for RCTs.

One could undermine assumption 2) by analysing the difficulty of approximating an 'ideal' RCT, as does Cartwright (2007). Most famously Heckman (1991) and more recently Worral (2007) provide wide-ranging systematic accounts of the various problems of 'internal validity' which affect RCTs. Furthermore, one could criticise RCTs on the basis that Bonell *et al.*'s (2012, p.2300) assertion that they 'provide minimally biased estimators or treatment effects' is misleading. As Deaton (2010, p.30) notes: 'RCTs are informative about the *mean* of the treatment effects, $Y_{i1} - Y_{i0}$, but do not identify other features of the distribution.' Really, then, RCTs only provide us with one estimator of each treatment effect. Policy makers might well be concerned with other estimators, for example median treatment effects and whether the distribution has 'fat tails' leading to subjects having outcomes that are very far from the mean.

To mitigate the concerns raised in the previous paragraph it can be argued that estimates of mean treatment effects, while not sufficient for policy making on their own, are a useful contribution to the scientific and policy-making effort. Therefore, this weakness of RCTs is acknowledged but not

considered damning for the purposes of this dissertation. Further, problems of internal validity are outside the scope of this dissertation; they have been addressed in detail by, for example, Banerjee and Duflo (2008) who argue persuasively that while many of them are real problems, they are better addressed by RCTs than by comparable methods which use non-random assignment. This dissertation supports the view of Bonell *et al.* (2012) that when evaluating the effects of development interventions at the micro level RCTs provide the highest levels of internal validity possible due to the power of randomization to isolate treatment effects from other potential causes. This is what Deaton (2010, p.28) calls 'the magic that is wrought by the randomization'. Nevertheless, it is the contention of this dissertation that the way in which RCTs of development interventions are currently being designed and implemented is impoverishing the evidence base. This is partly due to the problem of external validity, which is explained in the following section.

2.2 – The problem of external validity

The previous section demonstrates that the results of an RCT, as understood by researchers who adopt a successionist view of causality, can prove that the intervention T causes outcome O in test population φ with a given causal structure (Cartwright 2007). The great strength of the RCT when it comes to internal validity is that we do not have to specify anything about the causal structure of φ in order to conclude that T causes O because we know that as a result of randomisation any confounding factors can be assumed to act equally on different arms of the trial. However, this tells us nothing about other populations with other causal structures. If we consider a target population θ, then there is no guarantee that the causal structure of that population is not different in such a way as to mean that T will not cause O in θ. This is the problem of external validity: we need additional knowledge in order to extend our causal conclusions from the RCT beyond φ to other populations and such knowledge is difficult to generate. This section argues that a successionist account of causation cannot underpin any strategy for arriving at that knowledge.

Basu (2013) and Deaton (2010) have shown that the problem of external validity is not limited to populations 'external' to φ. Basu (2013, p.9) points out that we have no guarantee that the causal

structure of φ does not change through time in such a way as to undermine the claim 'T causes O in φ' and therefore 'tomorrow's population is a different one' requiring further premises to support the conclusions of the RCT. Deaton (2010) observes that we also have no guarantee that the causal structure of any non-random subset of φ is the same as that of φ. Thus, application of the insights of an RCT to any non-random subset of the test population also runs into the problem of external validity. Considerations like the above have led many authors considering evaluation based on a successionist account of causation to refer to a more general 'black box problem' in place of a problem of external validity (Astbury and Leeuw, 2010).

The standard response to the problem of external validity for RCTs of development interventions is to conduct replication studies (Pawson and Tilley 1997; Banerjee and Duflo 2008). In this strategy, interventions are conceptualised as 'products' that replication studies are used to 'accredit as effective' (Bonell *et al.* 2012, p.2300). A replication study involves conducting the same intervention in a different context in order to 'prove' that the differences between the test population and the target population are not a barrier to the causal efficacy of the intervention. The idea underpinning this approach is that '[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' (Banerjee and Duflo 2008, p.16). The key problem with this approach is that without a theory of how the intervention works in which causal structures there is no way of knowing in advance that the right 'covariates' have been chosen. 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 RCTs 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 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. It is therefore counterproductive to leave programme causation a 'black box', even if it appears easier and defensible from the point of view of internal validity

(Ravallion, 2009 p.38). A successionist account of causation is therefore insufficient and researchers must develop generative theories of intervention causation in order to rigorously argue from the findings of a particular RCT or set of RCTs to any statements about a target population θ.

Proponents of RCTs of development interventions who are sensitive to the weakness of the replication method but are not willing to surrender a successionist view of causation argue for the application of a concept such as Basu's (2013) 'reasoned intuition' when it comes to generalisation of findings from RCTs beyond their test populations. However, as Cartwright (2008, p.30) reminds us, an argument is only as strong as its' weakest premise. So, the argument that a particular intervention will have the desired effect on target population θ will only be as strong as our 'reasoned intuition' that θ is governed by a causal structure that is relevantly similar to that of test population φ. Unfortunately, 'reasoned intuition' seems to lead to very weak premises indeed. Basu (2013, pp.21-22) admits that it cannot be given a 'hard definition' but is equivalent to 'a leap of imagination' subjected to the scrutiny of reason by, for example, assessing its coherence with our existing beliefs. This seems to be an unnecessary retreat from rigorous social science to something that is not far from 'anything goes'. That such sacrifices are seen as acceptable by the same people who insist on extremely high levels of internal validity within trials is described by Cartwright (2007, p.19) as 'the vanity of rigour in RCTs'. The result of low attention to external validity is impoverishment of the evidence base. Pawson and Tilley (1997) argue that an excessive focus on evaluation of *whether* projects worked rather than *how* they worked held back thirty years of work in education, criminology and sociology. Heckman (1991) warns that economics may be falling into the same trap.

2.3 – 'Gold standard' thinking

The previous section warned that RCTs as currently implemented can lead to direct impoverishment of the evidence base resulting from the low external validity of their conclusions. This section argues that seeing RCTs as the 'gold standard' for evaluation of development interventions at the micro level also leads to an indirect impoverishment of the evidence base through the devalorisation of non-RCT evidence generation methods. If RCTs are seen as 'the gold standard', then other methods of evidence

generation must be seen as inferior (Cartwright, 2008; Deaton, 2010; Ravallion, 2009; Shaw, 1999). This means funders and researchers focus attention away from interventions which cannot be evaluated through RCTs, losing sight of important historical, institutional and structural questions and failing to generate theories of change. As Ravallion (2009 p.33) memorably warned, 'randomization is only feasible for a nonrandom subset of the interventions and settings relevant to development'. This is due, for example, to the political and ethical concerns that surround RCTs in some contexts as well as the practical difficulties involved in randomised allocation of subjects to different treatment arms (Heckman, 1991). The effect on qualitative approaches to evaluation is particularly marked as they are 'seen as failing to provide hard, reliable, factual data' (Sanderson, 2000 p.436).

Additionally, existing evidence generated through non-RCT methods is increasingly neglected as a result of the elevation of the RCT to 'gold standard' status. A good example of this process is the literature on cash transfers, systematic reviews of which increasingly ignore the excellent qualitative evaluations which were originally central to the literature but which are now rarely conducted. Three recent examples are DFID (2011), Fiszbein and Schady (2009) and Garcia and Saavedra (2013), none of which make any use of qualitative evidence despite the fact that they claim to be syntheses of 'current global evidence' (DFID, 2011 p.i). A historical look at the evidence base for cash transfer programmes in lower-middle income countries shows that qualitative evidence was key to the development of Mexico's PROGRESA/Oportunidades programme and Nicaragua's PRS programme which are now seen as best practice 'products' to be 'replicated' in low income countries (Adato *et al.*, 2010; Levy, 2006).

It is not just qualitative methods which are threatened by the adoption of the RCT as the 'gold standard'; the growing distrust of econometric analysis noted by Deaton (2010) has combined with the promise of simple answers from RCTs to create a climate in which 'development' is increasingly seen as a uniquely micro-level process that has very little or nothing to do with structural transformation of economies. Chang (2010) likens this to 'Hamlet without the Prince of Denmark' and bemoans the increasing slide towards '*ersatz* development'. While the ascent of the RCT as 'gold

standard' is not the only driver in this process, in the next section it will be argued that the promotion of realist RCTs could help to reverse this trend.

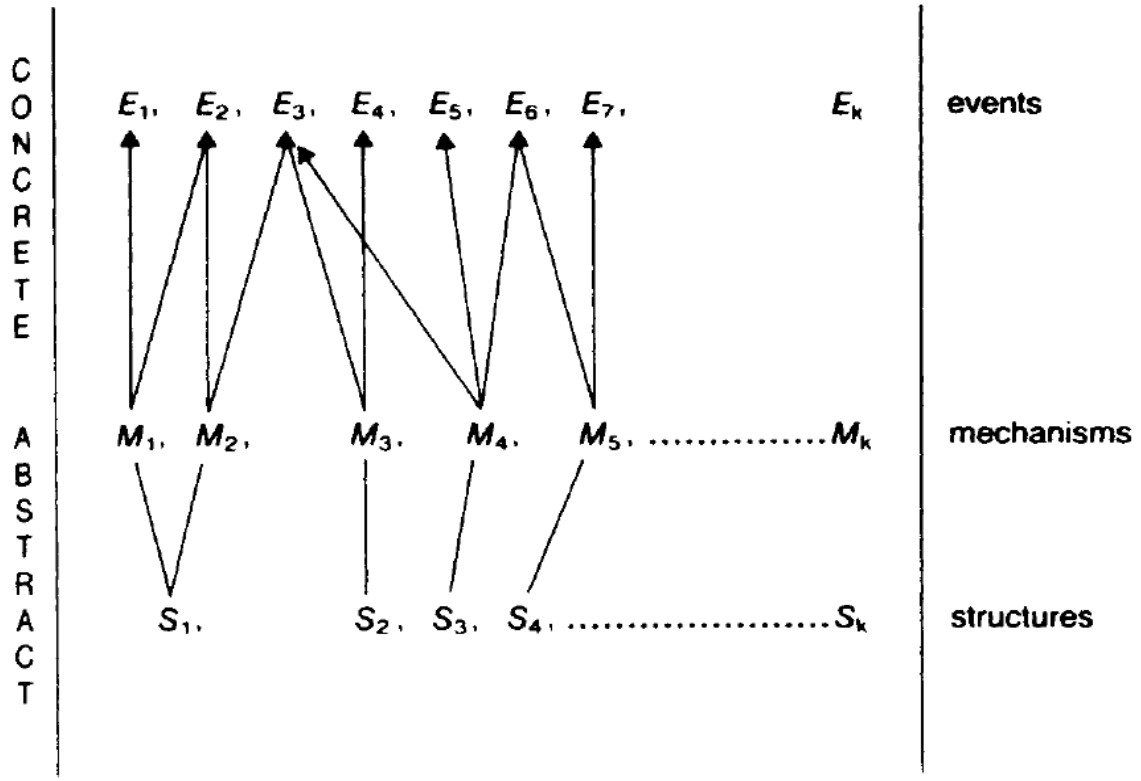
2.4 – Realist randomised controlled trials

This section examines what is meant by 'realist RCT' and argues that a move towards RCTs understood in this way would preserve the usefulness of the RCT method for attributing and measuring the effects of interventions whilst avoiding the problems of external validity and 'gold standard' thinking outlined in the previous two sections. Many successionist experimenters are 'too talented to be bound by their own methodological prescriptions' and admit that theories of causation have a role to play in the generalisation of the results of RCTs (Deaton 2010, p.4). For example, Banerjee and Duflo (2008 p.14) suggest that '[i]f we have a theory that tells us where the effects are likely to be different, we focus the extra experiments [replication studies] there'. However, in order to have 'a theory that tells us where the effects are likely to be different' researchers must move away from a successionist picture of causation and invest the effort required to develop a generative account of causation for the intervention in question. At this point, further experiments can shed light on much more than the 'covariation' of intervention effects with contextual factors and can be used to further develop the theory of not just 'what works' but 'for whom in what circumstances' (Pawson and Tilley 1997, p.220). Bonell *et al.* (2012) suggest that such trials could be seen as 'realist randomised controlled trials'.

Realism as understood in social science is an ontological framework 'through which the world is seen as an open system of dynamic structures, mechanisms and contexts that intricately influence the change phenomena that evaluations aim to capture' (Bonell *et al.*, 2012, p.2299). Specifically, realist ontology divides reality into three levels: events, mechanisms and structures (Sayer, 1992). Events constitute the level of reality to which we have the best access because we can directly observe them. For example, a person driving their car into a tree is an event. Events are generated by mechanisms (alone or in combination) such as inebriation or internal combustion. This generation may also be interfered with or even cancelled out by further mechanisms such as the action of an anti-lock braking

system. These mechanisms are the result of combinations of structures such as the human brain or the chemical composition of petrol. Across realists within social science, terminology is not completely standardised. For example, Connelly (2004, p.2) writes that '[r]ealism sees causal relationships as inhering in the specific nature of entities, so that each entity carries its own "causal powers and tendencies"'. Nevertheless, we can easily reconcile this statement with Sayer's ontology by interpreting 'causal powers and tendencies' as 'mechanisms' and 'the specific nature of entities' as 'the structure of entities'. Realists admit that we often cannot observe mechanisms or structures and must instead infer their existence from the observation of events (Bhaskar, 1975). Nevertheless, these levels of reality are taken to be sensible arenas for scientific study, albeit more 'abstract' than 'concrete' (Sayer, 1992 p.117). To illustrate realist ontology, Sayer's (ibid) figure 8 is reproduced below as figure 2.

Figure 2



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

It is outside the scope of this dissertation to argue for realist ontology from first principles. Fortunately, such an argument is not necessary. It was established in section 2.2 that a successionist account of causation provides no basis for the rigorous generalisation of results from RCTs. Nonetheless, we do have a strong intuition that some generalisation of RCT findings is possible. Harré (1970; 1972) has argued that a generative account of causation can provide a framework within which such work is possible and that this framework best describes the practices of our most successful science. Pawson and Tilley (1997, pp.61-63) agree, drawing on Huygens modelling of pendulum motion for examples of the experimental testing of hypotheses relating to causal powers. Given the necessity of a generative account of causation to defeat the problem of external validity, and realism's status as the ontology which underpins that account, realist ontology's *prima facie* reasonableness is sufficient for its adoption for the purposes of this dissertation. Realism's underlying ontology of structures, mechanisms and events and its epistemic strategy, retroduction, provide us with the means to rigorously interpret, generate and test generative causal theories (Bhaskar, 1975; Harré, 1970; 1972; Sayer, 1992).

Section 2.1 established that a successionist account of causation is one in which causation is a relationship between discrete objects which is reducible to their covariance. By contrast, the generative account of causation posits that causality 'concerns not a relationship between discrete events ('Cause and Effect'), but the "causal powers" or "liabilities" of objects or relations, or more generally their ways-of-acting or "mechanisms" (Sayer, 1992, p.104). As mechanisms can be counteracted by other mechanisms, causal claims are not reducible to regularities between separate objects or events; they are claims about how given mechanisms will interact in order to produce outcomes. When evaluating an intervention, Pawson and Tilley (1997) suggest distinguishing between the mechanisms through which the intervention is designed to act and those additional causal powers and liabilities (also 'mechanisms' in Sayer's tripartite ontology) which are present in any given context and which may facilitate or frustrate the intervention's causation of desired outcomes. The former are referred to as 'mechanisms' and the latter as 'context' with the basic unit of causal theorising being a 'context-mechanism-outcome combination' (ibid, p.220). The problem of external

validity for any study is resolved by reference to these context-mechanism-outcome (CMO) combinations, which provide a structure for the rigorous assessment of the similarities and differences between the causal structures of any given populations and the ways in which the mechanisms of the intervention can be predicted to bring about outcomes (or not) in those populations (Bonell *et al.*, 2012; Pawson and Tilley, 1997; Sanderson, 2000).

Pawson and Tilley have done the most to popularise realist ontology in the evaluation community but have been unfortunately restrictive in their methodological prescriptions. Since the publishing of *Realistic Evaluation* (ibid, 1997), 'realism' and 'the experimental method' have been conceptualised by many as necessarily in opposition. The most recent high-profile example of this is Westhorp (2014), an 'introduction to the key ideas in realist evaluation' designed for policy makers and published as a collaboration between the Overseas Development Institute (ODI), the Australian Department of Foreign Affairs and Trade (DFAT) and BetterEvaluation. In the guide Westhorp (ibid, p.7) construes realist evaluation as a set of design-level prescriptions rather than an ontological framework and associated epistemic strategy, and lists many situations in which 'realist evaluation can be more appropriate than experimental or quasi-experimental designs'. This follows in the tradition of Pawson and Tilley (1997; 2004) who argue that the use of RCTs reflects and entails a successionist conception of causation. Indeed, Pawson is a co-author of Marchal *et al.* (2013), which is a reply to the aforecited Bonell *et al.* (2012) and is entitled *Realist RCTs of complex interventions – An oxymoron*. Marchal *et al.* (2013, pp.124-125) state that RCTs are 'fundamentally built upon a positivist ontological and epistemic position', arguing that they fail to take into account the complexity of social causation by 'merely controlling for it'.

In reply to Marchal *et al.* (2013), Bonell *et al.* (2013, p.81) retort that 'methods don't make assumptions, researchers do'. This is an oversimplification of the relationship between method, epistemology and ontology in that they seem to suggest total independence between method on the one hand and epistemology underpinned by ontology on the other. This independence is the result of considering the connection between method and epistemology uniquely at the data interpretation stage

of the research process, as reflected in the authors' (ibid) assertion that 'Our methods don't determine our assumptions, we as researchers determine our understanding of the data deriving from these methods'. While this might be true, if we consider the design stage of the research process it is clear that some methods are not legitimate choices for realists. For example, persons of a Southern Baptist Christian ontological persuasion believe that the world is the creation of an omnipotent, omnibenevolent god and that the Bible, as the word of this god, is an infallible source. One legitimate research method for such a person is biblical exegesis. This research method is not legitimate for a realist social scientist because the retroductive research strategy requires an iterative movement from theory to observation of events and does not allow biblical exegesis as a legitimate source of confirmation. As no result of biblical exegesis would be admissible as confirmation or rejection of theory, there can be no expectation that it will be useful and it is therefore not a legitimate research method. This does not mean that the results of biblical exegesis cannot be interpreted by realists; they can be interpreted but will always be found to be irrelevant. The fact, then, that a method's results are interpretable within a given epistemic strategy does not mean that the method is a legitimate choice for adherents to that strategy when engaged in research design. What is needed for a method to be a legitimate choice is an account of how the data generated by the method can be interpreted in such a way as to help answer research questions. Bonell *et al.* (2012) have provided the beginnings of such an account based on the observation that RCTs can be used to test causal theories rather than merely to establish covariation. This involves the explicit theorisation of the intervention mechanisms and the contextual factors necessary to bring about desired outcomes. The specific design choices required to facilitate this process will be explored in section 3.2.

Realist epistemology appears to offer a way of salvaging the value of RCTs in social science without leading to impoverishment of the evidence base. Realist RCTs require a theory of 'how the intervention works, for whom under what circumstances' thereby allowing an argument to be developed which rigorously underpins the assumption that causal structures in a target population will be relevantly similar to those in the test population. This avoids the problem of external validity which plagues successionist RCTs as outlined in section 2.2. By abandoning a restrictive successionist

account of causation, social science realism facilitates the realisation that 'there is no gold standard' when it comes to evidence generation. RCTs are not seen as a 'gold standard' because other methods are also desirable or even required in order to generate and evaluate causal hypotheses about CMO combinations. Rather, RCTs are seen as one very useful tool in the researcher's toolkit, avoiding the problems raised in section 2.3.

Requiring explicit specification of generative programme theory strengthens the RCT method against another criticism often raised against it: that 'approaches founded upon the assumptions of stability and equilibrium, of linearity in the relationship between variables, and of proportionality of change in response to causal influences…are not appropriate in seeking to understand social systems that exhibit complexity' (Sanderson, 2000, p.442). By construing the social world as a series of open systems, realist ontology makes room for instability, 'dissipative systems' or nonlinear phenomena such as threshold effects. All the consequences of complexity can be borne in mind when interpreting the results of RCTs as part of a retroductive epistemic strategy. It is the insistence on a successionist notion of causality in which covariance equals causation which is vulnerable to Sanderson's criticism, not any particular method of generating and measuring outcomes.

2.5 – Critical realist randomised controlled trials

Porter and O'Halloran's 2012 paper does not go as far as Bonell *et al.*'s paper from the same year in that it argues for the concurrent use of realistic evaluation and positivist RCTs rather than the full integration of the RCT within a realist approach. However, in one respect Porter and O'Halloran (2012, p.18) apply realist epistemology more thoroughly than Bonell *et al.*: they criticise realistic evaluation because 'it rejects the critical turn of Bhaskar's realism' and 'replicates the technocratic tendencies inherent in evidence-based practice'. Section 3.2 will draw on Cartwright's (2008) argument that interventions are underpinned by implicit causal models even when they appear to be 'theory neutral'. Similarly, it is important to remember that 'claims to knowledge are claims to power' and that 'technocratic' solutions can conceal the promotion of the interests of the powerful (Demeritt, 1996, p.485).

Pawson and Tilley (2004, p.2) use a lot of ethically-charged language in the introduction to their 2004 book chapter, calling interventions 'hypothesis about social betterment' and the means by which 'wrongs might be put to rights, deficiencies of behaviour corrected, inequalities of condition alleviated'. However, the promise of this value-laden introduction is betrayed by an exclusively technocratic treatment of evaluation. For example, the authors repeatedly reference realist evaluations of closed-circuit television (CCTV) surveillance systems such as Tilley (1993), Painter and Tilley (1999) and Gill (2003) without acknowledging that the use of CCTV raises ethical as well as technical questions. Similarly, Bonell *et al.* (2013, p.81) say that they adopt 'a critical but realist ontological and epistemological perspective', hinting at an awareness of the influence of critical theory on Bhaskar and Sayer's work. This awareness is not reflected in their methodological guidelines, however; as no mention is made of the need to take seriously both the emancipatory and the oppressive potential of social science (ibid, 2012; 2013).

Porter and O'Halloran (2012), by contrast, provide examples from their own work in complex public health interventions of the need to examine the desirability of the outcomes of an intervention as well as the efficacy of the context-mechanism combinations employed for their achievement. The authors (ibid, p.24) conclude that 'without a utopian vision that consciously and constantly maintains the criteria of improvement of clients' physical, psychological and social needs as its core goal, then the danger of diverting away from that goal into bureaucratically-driven research will be ever-present'. It is beyond the scope of this dissertation to analyse the particular forms of utopian thinking which have been suggested as philosophical underpinnings for a normative ethics of critical realism. However, it is clear that outcomes of interest for any intervention are always finally justified with reference to some value judgement and that this should be rendered explicit in any evaluation of that intervention. For example, interventions which seek to reduce poverty do so because poverty is judged to be a bad thing for people. Sayer (1992, p.39) is careful to emphasise that it is the duty of social scientists to examine 'common sense' notions such as 'poverty is bad' in order to uncover potential sources of error or complexities which might be hidden within expedient everyday discourse. As Sayer (ibid) argues, 'science is redundant if it fails to go beyond a common-sense understanding of the world'.

Therefore it is incumbent upon researchers to have a more precise conception of, for example, why poverty is 'bad'. The World Bank's *Handbook on Poverty and Inequality* (2009, p.1) defines poverty as 'pronounced deprivation in well-being'. This common-sense concept can itself be further unpacked; Sen (1999) famously argues that well-being is the capability to function as a human being in society. If an intervention sought to reduce 'poverty' by boosting household income and this outcome were implicitly justified through plausible interpretations of common-sense concepts such as those above, it would behove the evaluator(s) to ask 'is there any *prima facie* reason to suspect that this intervention might reduce net well-being in subjects despite boosting household income?' If so, the evaluation design clearly ought to include some way of measuring other proxies for well-being in order to eliminate this doubt. The theoretical literature thus provides a strong argument that realist RCTs ought to incorporate the critical element of Sayer and Bhaskar's critical realism in order to guard against perverse consequences or bureaucratic oppression through exclusively technocratic approaches to development. Rather than referring to 'critical realist RCTs' this dissertation will continue to refer to 'realist RCTs' as this terminology has become standard in the literature. However, this realism should be interpreted to include a critical approach to the moral frameworks implied by interventions and more generally to the role of the researcher.

3 – FROM THEORY TO PRACTICE

3.1 – Overcoming inertia

As the preceding chapter has established, the theoretical basis for 'realist RCTs' is strong. However, this is not sufficient for an impact on research practice. This can be seen, for example, in the complex public health intervention literature. The 2000 Medical Research Council 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'. This section outlines some mechanisms which lead to this kind of inertia and suggests a strategy to overcome them.

Sayer (1992, p.254) understands that people who work within 'routinized practices and their associated ideas' have limited freedom to change their way of thinking. However, he suggests that there is a categorical difference between 'researchers and the researched' in that researchers can be considered to be 'primarily leading a life of reflection' within which it is easy to change one's ideas. It is more plausible to suggest that, while there is clearly a difference of degree, there is no categorical difference between the situations of the researcher and the researched. Social scientists are caught up in 'routinized practices' which make it extremely difficult for them to interpret and to take seriously challenges to the ontological, epistemic and methodological commitments which are implicit within their knowledge community. In practice, this means epistemology and ontology remain unexamined most of the time within most disciplines and researchers fall unthinkingly into copying the methodologies of exemplar studies within their discipline (Kuhn, 1969). Support for this interpretation can be found in the interdisciplinarity literature where concerns over the difficulty of acknowledging underlying assumptions from within disciplinary thinking lead researchers like Harriss (2002, p.2) to warn that disciplines need to be 'saved from themselves' through the application of 'anti-discipline'. For Harriss (ibid) anti-discipline usually comes from cross-disciplinary work. However, as Lélé and Norgaard (2005, p.972) have persuasively argued, 'the structure of scientific

knowledge and the differences in epistemologies, theories, and methods among scientists have little to do with what have historically been called disciplines'. For this reason, they urge us to 'forget disciplines; think scientific communities'.

Thinking about 'scientific communities' facilitates an understanding of why so much inertia surrounds the reform of the excessively successionist approach to the evaluation of development interventions at the micro level. The scientific community which implicitly or explicitly adopts a successionist view of causation is extremely extensive, meaning that it is difficult to expose its members to 'anti-discipline' which would prompt them to examine and perhaps even revise their successionist assumptions. This difficulty has been heightened by the propensity for realists to continue to criticise others for their 'positivist' assumptions despite the fact that it is widely acknowledged that '"positivism" is now a term … of abuse rather than illumination in the social sciences' (Oakley, 2000). Because this is the case, very few social scientists identify as positivists. As Williams (1983, p.239) memorably remarked, 'positivism' has become 'a swear-word by which nobody is swearing'. It is for this reason that this dissertation has been careful to avoid the term 'positivism' wherever possible, preferring to talk specifically about the successionist account of causation as the deficient aspect of positivist thinking. Because very few researchers identify as positivists, a realist understanding of RCTs will not be promoted by attacking 'positivist RCTs'. Therefore, this dissertation aims to engage with successionist experimenters in the way that they see themselves, as pragmatists. This is achieved by critically appraising high-profile nascent exemplar trials within the mainstream, successionist-inflected literature from a realist perspective. This critique aims to disrupt the process of methodological reproduction discussed in the previous paragraph by introducing challenging anti-discipline to these high-profile trials, encouraging those researchers who would take them as exemplars to examine their implicit ontological and epistemic underpinnings.

3.2 – Realist randomised controlled trials at the design level

Before critiquing the exemplar trials analysed in chapter four, it is necessary to interpret the design level consequences of the ontological position and epistemic framework argued for in chapter two.

The key insight of section 2.4 was that RCTs should be based on a theory of 'how the intervention works, for whom under what circumstances'. The most important design-level consequence of this insight is that this theory must be explicitly specified. As Cartwright (2008) reminds us, every programme evaluation is based on a theory of intervention causation whether this is acknowledged or not. Sanderson (2000, p.437) agrees, lamenting that not explicitly referring to this theory leads to 'an implicit adoption of prevailing, taken-for-granted theoretical frameworks which … undermine the capacity of evaluation to produce knowledge which is useful in informing social action.' The causal theories which underpin a realist evaluation therefore need to be explicit throughout the research design process from what Mayoux (2006, p.124) systematises as the 'scoping' stage and must be explicitly present in all disseminated materials. Section 2.5 argued that the normative framework necessary to motivate the intervention is also highly relevant for researchers and policy-makers. It should therefore also be rendered explicit at all stages of the research process.

A realist RCT, then, must explicitly specify a theory of 'how the intervention works, for whom under what circumstances'. The question remains: what form should this theory take and how extensive does it have to be? Bonell *et al.* (2012, p.2304), following Pawson and Tilley (1997) argue that these theories should be 'mid-level theories that … aim to explain how context and an intervention's underlying mechanisms interact to produce outcomes'. Cartwright assess how extensive this theory has to be in order to be sufficient for policy purposes in her 2008 working paper, arguing that nothing short of a full causal model of the test population will be sufficient for policy-making and further theory-building purposes. Cartwright's minimum acceptable specification of this model follows:

> *'1. A list of the causes relevant to the targeted effect that will operate in the target situation. This includes*
>
> *1.a. the causes present in the situation independent of the policy action*
>
> *1.b. any changes in this set of causes introduced in implementing the policy.*

2. *A rule of combination that calculates what should happen vis-à-vis the targeted effect when those causes operate together.'*

(Cartwright, 2008, p.9)

Although this seems arduous, Cartwright (ibid) argues that in the absence of an explicit formulation of such a model 'when one bets on an effectiveness counterfactual, one is betting, willy-nilly, on the causal model that underwrites it'. She continues that '[t]he whole point of evidence-based policy is that bets like this should be taken consciously and be as well informed by evidence as is practicable' (ibid, p.41). This argument seems sound from a realist perspective. An unwillingness to engage in such arduous and error-prone theory building might motivate the adoption of a successionist account of causation, but it has been demonstrated that the attractive freedom from theorising offered by this way of thinking is an illusion and leads to insurmountable problems of external validity. Therefore, realist RCTs must be accompanied by causal models of the test population which satisfy Cartwright's specification, allowing researchers to 'describe the theory of change in evaluation reporting as well as exactly how the given study aims to examine the theory of change' (Bonell *et al.*, 2012, p.2304).

One implication of having a causal model which meets Cartwright's (2008) conditions is that it should be possible to identify intermediate outcomes which will be achieved on the way to the primary outcomes of interest. These 'pathway variables' can be measured in order to test the theory behind the intervention (Bonell *et al.*, 2012, p.2303). These outcomes may also be desirable in their own right, and can be identified as 'secondary outcomes'. For example, Winkleby, Feighery *et al.* (2004, cited in Bonell *et al.*, 2012) measured and analysed 'self-efficacy' as both a secondary outcome and a pathway variable to 'tobacco-avoiding behaviour'.

In order to 'examine the theory of change' rather than merely attempting to 'accredit' a specific 'intervention product', realist RCTs must evaluate the different components of complex interventions separately as well as in combination (Bonell *et al.*, 2012, p.2303). Fortunately, this is already common

practice within the complex public health literature; what Banerjee and Duflo (2008, p.6) call 'multiple treatment experiments' are well understood as was highlighted in section 2.1.

Bonell *et al.* (ibid, p.2303) also suggest that there should be 'a more strategic, coordinated approach to testing the effects of interventions and their components in different contexts using consistent measures where possible'. This suggestion is certainly warranted and indeed more coordination is already being called for by mainstream, successionist researchers (Banerjee and Duflo, 2008; Wong *et al.*, 2013). However, 'consistent measures' are clearly not enough to facilitate sustained theoretical integration between diverse researchers. Bonell *et al.* (2012) suggest that the development and testing of hypotheses about 'context-mechanism-outcome' combinations will suffice. However, this is a misreading of Pawson and Tilley (1997, p.220, emphasis in original) who call not just for hypothesis testing but for the development of '*families* of theories specifying *typologies* of successful context-mechanism-outcome combinations'. These families of theories can be refined as different researchers move retroductively between the theoretical level of 'abstraction' and the 'concrete' level of observation via the research design level of 'specification' (ibid, p.121; Sayer, 1992, p.87). The use of typologies in realist social science has a long history as 'a potentially useful methodological tool providing a vital link between theory and practice' (Whatmore *et al.*, 1987, p.22; Allen and McDowell 1989). Typologies of CMO combinations could provide a means of linking not only theory and practice, but also researchers, even those from different scientific communities. The way in which they could do this is by constituting '*boundary concepts* that allow ... conceptual communication' across 'boundaries' within social science (Mollinga, 2010, p.S-4). Boundary concepts are terms with the same referent but which capture different meanings of that referent depending on the community in which they are used. As Mollinga (ibid) puts it, they are 'different abstractions from the same "thing"'. Typologies of CMO combinations clearly fit this definition. By contributing to the formation and retroductive appraisal of CMO typologies, realist RCTs could provide a theoretical framework which allowed for a more coordinated approach to evaluation of development interventions by facilitating inter- and trans-disciplinary work.

Section 2.4 argues that a realist understanding of RCTs avoids the problems raised in section 2.3 by conceptualising RCTs as a useful tool at the researcher's disposal rather than a 'gold standard'. This opens the door for the embedding of RCTs within a mixed-methods research strategy, allowing for 'methodological triangulation' as argued for by Yeung (1997). There are many advantages of methodological triangulation for internal validity. For example, Devereux *et al.* (2006) describe a mixed-methods evaluation of a cash and food transfer programme in which quantitative and qualitative surveying of subjects indicated no increase in spending on temptation goods. However, the use of indirect probing in focus groups to ask some subjects about the spending habits of other subjects revealed widespread, credible complaints of increased spending on temptation goods with wives in particular complaining of husbands 'diverting some cash to other wives and girlfriends and neglecting their children' or 'frittering away' money on 'beverages' (ibid, p.8). A mixed methods approach also benefits external validity. For example, by allowing researchers to use detailed case analysis, qualitative data can also be generated about intervention causation, informing theory and potentially allowing for quantitative assessment in later trials. As Mayoux (2006) observes, qualitative, quantitative and participatory methods all have strengths and weaknesses and may have their roles to play in the long process between initial engagement with a research question and the dissemination of results.

In summary, realist RCTs should:

1. Develop an explicit causal model of the outcomes of interest within the test population to facilitate:
    a. An explicit account of intervention causation
    b. Testing of that account.
2. Measure pathway variables to inform causal theory
3. Evaluate the different components of complex interventions separately as well as in combination
4. Contribute to the formation and iterative reappraisal of CMO typologies

5. Be incorporated into a mixed-methods approach to data generation
6. Explicitly outline the normative framework from which the intervention draws its justification

4 – CASE STUDY ANALYSIS

4.1 – Sampling rationale

Chapter two has established the theoretical case for realist RCTs and chapter three has argued that in order to change practice it is necessary to critique the exemplar studies within a scientific community. To that end, this chapter presents a critique of two exemplar trials of development interventions at the micro level. In an effort to maximise the chance that change might be achieved through this research, the trials have been purposively sampled to facilitate this. It is not pretended that the sampling method adopted was random, nor that the sample is representative of the wider literature – the sample was selected in order to facilitate change of the literature, not to describe it. The sampled trials are from the rapidly-growing lower income country cash transfer literature. This is because, borrowing military metaphors from Klein (1990, pp.77-79) and as section 2.3 has outlined, this literature has become a 'battlefield of ideas', and one on which successionist thinking seems to have 'gained a lot of ground'. The sampled trials are both reported by papers that were disseminated within the last three years. This is due to the fact that the science of evaluation is developing rapidly and it was deemed desirous to critique 'cutting edge', 'best practice' trials (Wong *et al.*, 2013). The trials have been purposively sampled to be as high-profile as possible. Case study one was sampled for its high profile in the academic literature as determined through citation analysis. Number of citations was adjusted for the length of time elapsed since publication following a keyword search on Google Scholar for 'randomised randomized controlled trial cash transfer' with automated filtering for date and manual filtering for subject area. Google Scholar citation analysis was used rather than Scorpus or ISI measures because this best captures citations within both the natural and social sciences as traditionally construed (Harzing, 2010). Case study two was sampled for its high profile in policy circles. Despite being unpublished, the trial results have gained a huge amount of attention in policy circles via widespread reporting in the popular press (i.e. The Economist 2013; Goldstien 2013; Karnofsy 2013; Kestenbaum 2013). In previous work submitted for this degree, I proposed the use of qualitative and participatory methods to triangulate the results of this trial. This was partly in order to address suspicions that the use of self-reported consumption measures to assess spending on spending

on temptation goods was seriously flawed. My concerns in this previous work were therefore with the trial's internal validity. The analysis offered in this dissertation is very different, albeit complementary.

The papers reporting the sampled trials follow:

1.
    a. Baird, S.J., Garfein, R.S., McIntosh, C.T. and Özler, B. (2012), "Effect of a cash transfer programme for schooling on prevalence of HIV and herpes simplex type 2 in Malawi: a cluster randomised trial", *The Lancet*, Vol. 379 No. 9823, pp. 1320–1329.
2.
    a. Haushofer, J. and Shapiro, J. (2013a), *Welfare Effects of Unconditonal Cash Transfers: Pre-Analysis Plan* (Unpublished Working Paper).
    b. Haushofer, J. and Shapiro, J. (2013b), *Policy Brief: Impacts of Unconditional Cash Transfers* (Unpublished Working Paper).
    c. Haushofer, J. and Shapiro, J. (2013c), *Household Response to Income Changes: Evidence from an Unconditional Cash Transfer Program in Kenya* (Unpublished Working Paper).

4.2 – Case study one: Baird *et al.* (2012)

4.2.1 – Description

Published in the *Lancet* Feb 2012 and cited 134 times to date (Google Scholar, 11/09/2014), this paper reports a cluster-randomised controlled trial, which 'assessed the efficacy of a cash transfer programme to reduce the risk of sexually transmitted infections in young women' in Zomba district, Malawi (Baird *et al.*, 2012). Households were randomly assigned to treatment arms as outlined in section 2.1 and Figure 1. By identifying the trial as an *efficacy* trial, the authors suggest that they intend to assess the modest claim that such an intervention can, under favourable conditions, produce

a statistically significant result (Singal *et al.*, 2014). They are therefore primarily concerned with internal validity.

4.2.2 – Causal model

Despite being primarily concerned with internal validity, the authors begin to elaborate a very general causal model in order to justify the *a priori* reasonableness of the intervention and lay the groundwork for the cumulation of knowledge through later *effectiveness* trials. This causal model is thin and does not meet Cartwright's (2008) standards outlined in section 3.2. The authors observe that women and girls in sub-Saharan Africa (SSA) are disproportionately at risk of HIV infection and that 'lack of education and an economic dependence on men are often suggested as important risk factors' along with 'poverty' (Baird *et al.*, 2012, p.1320). They go on to observe that '[c]onditional cash transfer programmes aim to reduce current poverty and, by investment in the education of children, future poverty' and then argue that '[b]ecause conditional cash transfer programmes increase household income and school enrolment, they are particularly suitable for investigation of the importance of education and poverty as risk factors for HIV' (ibid, p.1321).

While this causal sketch is situated in SSA, no mention of causally-relevant contextual factors is made apart from the observation that 'Zomba district… is characterised by poverty, low school enrolment, and a high prevalence of HIV' (p.1321). Even on the very minimal causal model outlined, there are other important contextual factors which have not been addressed. For instance, 'economic dependence on men' is one of the three risk factors identified above, but no analysis of its prevalence in Zomba district is offered. In order to demonstrate the efficacy of the intervention in favourable circumstances as defined by this causal model, the trial should at minimum establish the presence of this 'risk factor' in the context.

4.2.3 – Testing of the causal model

The trial makes no attempt to test the underlying causal model. Enumeration areas were sampled using stratified random sampling in order to represent urban, near rural and far rural areas but the only

rationale for this is that ‘block randomisation by geographical stratum helps to prevent bias that might be caused by differences in the underlying prevalence of HIV’ (p.1327). This ‘block randomization’ offered an opportunity to elaborate contextual factors relevant to any heterogeneity of effects by type of area and thereby test the underlying causal model, but this opportunity was not taken.

No intermediate outcomes were tested. The authors generated data through surveys that suggested ‘transactional sex’ was common among subjects and it is observed that intervention recipients chose younger partners whereas the control group were more likely to have sex with partners who could aid them financially. However, transfers were made to heads of households. It is unclear whether the intervention was supposed to have increased the disposable income of girls themselves but, if this is implicitly a channel through which the intervention was supposed to have operated, then this could have been tested. For example, girls could have been surveyed on their level of disposable income or feelings of economic empowerment or some other proxy determined to be appropriate through piloting.

The most striking omission regarding testing of the underlying causal theory is that the study ‘was not powered to detect… heterogeneity of effects between the conditional and unconditional cash transfer programme groups’ (ibid, pp.1323-1324). Thus, while the rationale for the intervention exclusively mentions ‘conditional cash transfer programmes’, the tested intervention is, in effect, a combination of conditional and unconditional cash transfer programmes, clouding the attribution of effects without providing the statistical power to detect differences between these forms of the intervention.

4.2.4 – Mixed methods integration

This trial does make use of mixed methods in that it triangulates biological data with survey data to improve the internal validity of findings. However, the paper still demonstrates a degree of the ‘gold standard’ thinking critiqued in section 2.3. On page 1320, the authors state: ‘although poverty, especially poverty of women, has been suggested as a major risk factor for HIV, evidence is mixed. We are aware of only one cluster randomised trial with biological outcomes of a structural

intervention to prevent HIV in women.' This section judges non-RCT evidence to be admissible, in that it discusses the 'mixed' evidence-base. However, the RCT is considered the gold standard. This is also reflected in the 'systematic review' process related in panel two, which sets out a categorical preference for RCT evidence. A willingness to take qualitative data seriously could have aided to create a more comprehensive causal model as well as offering the possibility of mixed-methods triangulation of the pathway variables mentioned in the previous subsection.

4.2.5 – Normative framework

No normative framework is elaborated to justify this intervention. It is assumed that HIV is a 'very bad thing' that considerable effort should be expended to minimise, and this assumption seems reasonable. However, even in this clear case of common sense, Sayer's (1992, p.39) warning that 'science is redundant if it fails to go beyond a common-sense understanding of the world' is pertinent. There are *prima facie* reasons to suspect that cash transfers funded by a foreign organisation can have negative consequences in the right contextual environment, undermining local institutions, particularly in highly aid-dependent countries (Rajan and Subramanian, 2007). While it may be outside of the scope of an efficacy study to test for such effects, it could be considered incumbent on the authors to engage with this debate. A critical realist engagement with the active ethical questions could have rendered the intervention more attractive to policy makers, for example at the IMF, who share Rajan and Subramanian's (ibid) concerns.

4.2.6 – External validity and prospects for cumulative contribution

The paper concludes very modestly that 'an intervention without direct focus on sexual behaviour change can lead to meaningful reductions in HIV and HSV-2 infections'. While this appears justified on the basis of the internal validity of the trial, two considerations weigh against the validity of the conclusion. The first is the authors' own observation that of the 'few' other trials that 'have assessed behavioural interventions with biological outcomes', 'none have recorded a significant effect on HIV'. This should suggest a lower level of confidence in the results when taken as part of the bigger picture, but no suggestion is made that the authors have adjusted their confidence accordingly.

Further, the use of the word ‘meaningful’ reflects confusion in the paper between the concepts ‘significant’ and ‘statistically significant’. ‘Significant’ is used by the authors throughout the ‘results’ section, where presumably what is meant is ‘statistically significant’. The earlier use of ‘significant’ makes the jump to the use of ‘meaningful’ seem natural, but a separate argument is required to establish what is meant by a ‘meaningful’ reduction in HIV infection in this population if ‘meaningful’ has some meaning over and above ‘statistically significant’. Nevertheless, the study does seem to prove that a cash transfer programme *can* reduce the rate of HIV infection in a given population.

The question remains, ‘what use is this trial to the wider research and policy community?’ While the study has proven that a cash transfer programme *can* reduce the rate of HIV infection in a given population, this was reasonable to conclude *a priori*, as the trial’s own authors suggest in their introduction. Researchers and particularly policy makers require the answer to a much more demanding question than the question answered by the trial: they need to know whether such an intervention is likely to reduce the rate of HIV infection in a given target population, requiring an argument for external validity of the trial’s findings. The trial cannot provide premises for such an argument because it fails to develop a causal model of the test population which is sufficiently detailed to underpin an assessment of whether the causal structure of a given target population is relevantly analogous to that of the test population, as argued in section 4.2.2. The paper begins to argue that the test population is representative of Malawi, observing that ‘[t]he prevalence of HIV (3·7%) in unmarried women younger than 25 years in southern Malawi was nearly identical to the overall HIV prevalence (3·8%) in the control group’ and arguing that ‘[t]herefore, our sample of school-aged girls seems to be representative of the population in the study area’. However, the similarity observed is only one of outcomes and not of causal structure. A realist attention to the causal features of the test population could have motivated an argument that the test population was representative of Malawi, but similarity of outcomes is not enough.

Despite defining the trial as an 'efficacy trial' and arguing for the modest conclusion that a cash-transfer intervention *can* reduce HIV infection rate, the authors make some policy suggestions which indicate that they believe the findings of the trial have some application to other contexts. They suggest that their results 'indicate that cash transfer programmes could be attractive to policy makers in sub-Saharan Africa when they consider the full array of benefits that they might provide', and even that 'costs for a scaled up cash transfer programme would yield a cost of only $5000 per HIV infection averted' (p.1328). From a realist perspective, it is too early in the evaluation of the effects of cash transfers on HIV infection rates to make such suggestions. In order to lend some support to such assertions, this trial would have had to develop a more explicit model of intervention causation, to have tested that model, to have thereby contributed to a typology of CMO combinations which provide an account of what works, for whom, in what circumstances, and to have engaged with the wider debate about the desirability of such an intervention. These are the benefits which could have resulted from designing this trial as a realist RCT.

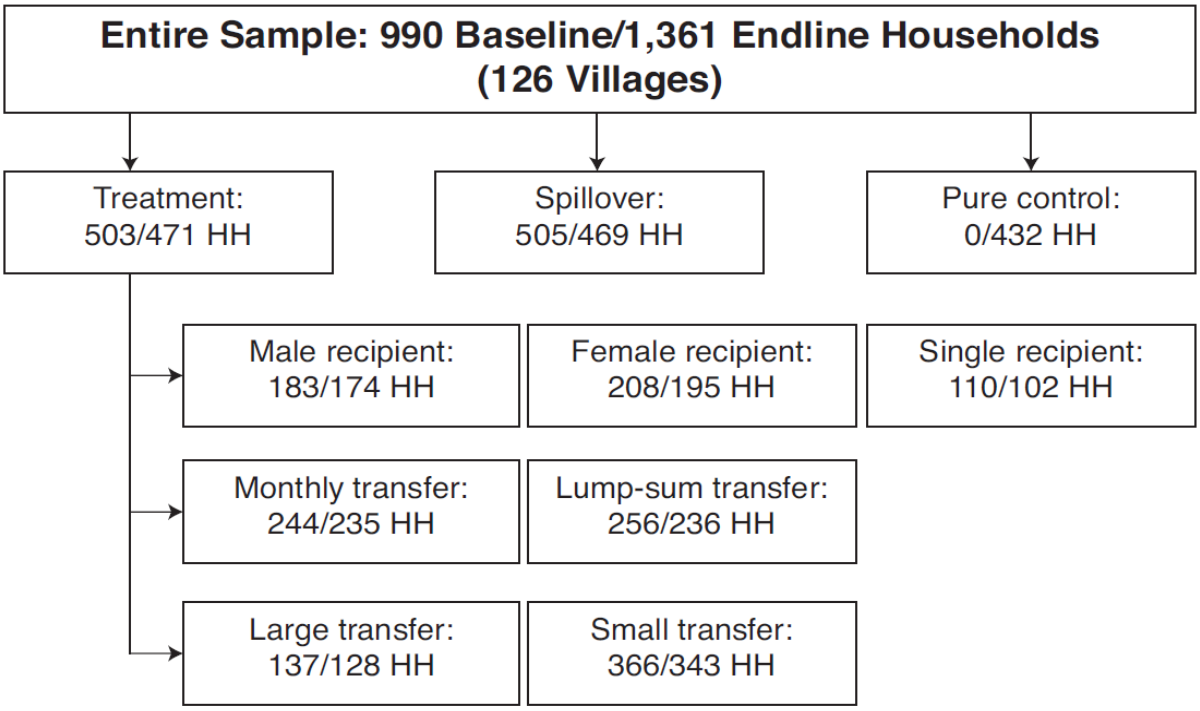
4.3 – Case study two: Haushofer and Shapiro (2013a; 2013b; 2013c)

4.3.1 – Description

This section draws on three 'unpublished' but very widely consulted working papers which report on a cluster-randomised controlled trial of an unconditional cash transfer to households in Rarieda, Kenya administered by the INGO GiveDirectly and evaluated by Haushofer and Shapiro under the auspices of Innovations for Poverty Action. The three papers consist of a) a pre-analysis plan posted to socialscienceregistry.org (2013a), b) a 'policy-brief' (2013b) and c) a paper (2013c) aimed at contributing to the income-change response literature in economics. The pre-analysis plan does not report any results, being designed as an instrument to constrain the authors in their data analysis 'to prevent data-mining and cherry-picking of results' (Haushofer and Shapiro, 2013b, p.9). The policy brief and economics paper report the same set of results with different sets of interpretations to accomplish different objectives: one aims to inform policy-makers and the other aims to contribute to theoretical debates in economics. The intervention itself involved the transfer via mobile money system M-Pesa of either large or small lump sums or monthly instalments to treatment households

cluster-randomised into treatment and control at the household and village level as specified in the authors' figure 1, reproduced here as figure 3.

Figure 3



Notes: Diagram of treatment arms. Numbers designate baseline/endline number of households in each treatment arm.

(Haushofer and Shapiro, 2013b, p.25)

4.3.2 – Causal model

No causal model was explicitly specified in the policy brief, but the economics paper draws on a large theoretical literature to elaborate an abstract, largely context-independent theory of causation for the outcomes of interest. The contextual factors that do feature in this theory are highly abstract concepts such as the property of being 'savings-constrained', either physically, socially or behaviourally. No analysis of the causal structure of the target population in particular is conducted beyond the observation that the average monthly household income is around 200 USD and that recipients are 'poor households in Kenya'.

4.3.3 – Testing of the causal model

Random allocation to the different arms of the trial depicted in figure 3 and careful attention to the power of the trial enabled the authors to test theories relating to 'three design features of unconditional cash transfers: whether the transfer recipient is the husband or the wife within the household, whether the transfer was made in a single a lump sum, or in nine monthly instalments, and the size of the transfer' (ibid, p.2). However, in the absence of a model of the causal structure of the population, discussion of the policy and theoretical implications of results runs into serious problems of external validity as will be discussed in section 4.3.6.

4.3.4 – Mixed methods integration

There was no integration of this RCT into a mixed-methods research programme. Such an integration would have offered possibilities to improve internal validity as discussed in my previous work and section 4.1. It would also have facilitated the elaboration of a tentative, testable model of the causal structure of the population in order to facilitate external validity for findings.

4.3.5 – Normative framework

Identically to case study one, no normative framework is elaborated to justify this intervention. The same considerations discussed in section 4.2.5 therefore also apply to this trial with a consequent reduction in the appeal of this trial to policy makers compared to a counterfactual realist RCT.

4.3.6 – External validity and prospects for cumulative contribution

Thanks to the highly abstract nature of those contextual factors which are mentioned by the authors, and their having omitted to employ qualitative methods to further specify these factors, attempts to contribute to the wider literature run into problems of external validity. The policy brief provides an outstanding example of the 'vanity of rigour' of successionist RCTs discussed in section 2.2 following Cartwright (2007, p.19). It is entitled 'Policy Brief: Impacts of Unconditional Cash Transfers' rather than 'Policy Brief: Impacts of *an* Unconditional Cash Transfer'. The authors go on to claim that 'Transfers', implicitly *in general*, 'allow poor households to build assets' and 'increase consumption'

and ‘reduce hunger’ and ‘do not increase spending on alcohol and tobacco’ etc. (Haushofer and Shapiro, 2013b, p.2). Regardless of any possible problems of internal validity, it is clear that such general conclusions cannot be supported from a realist perspective. Even if the authors believe that context is irrelevant to the action of the causal mechanisms which produced these outcomes in the test population, no argument has been provided to justify this belief. Presumably these findings are intended to implicitly have some limited area of application as the authors earlier mention cash transfers’ increasing profile as ‘a potential alternative poverty alleviation strategy’ (ibid, p.1). This, coupled with the policy recommendation that ‘transfers reduce hunger’ and the targeting strategy of GiveDirectly suggest that the policy recommendations should be understood as applying to transfers to the extremely poor. Even granted this implicit limitation, many questions remain. For example: ‘how poor should recipients be in order for these outcomes to be expected?’ Haushofer and Shapiro’s trial has clearly not provided the answer, though realist RCTs might be able to by constructing typologies of CMO combinations.

Despite a higher level of theoretical engagement, similar problems affect the economics paper. An instructive example is the discussion of spending on ‘temptation goods’. Haushofer and Shapiro (2013c, p.35) state that the trial was designed ‘to answer several longstanding questions in economics’ including ‘how do households respond to income changes?’ On this question, the authors consider the trial finding that ‘[a]lcohol and tobacco expenditures did not increase’ (ibid) and conclude that ‘simple cash transfers may not have the perverse effects that some policymakers feel they would have’. Setting aside any potential problems of internal validity, it is not clear that the trial has contributed any understanding to questions about spending on temptation goods *in general*. As the authors note in their introduction, there is ‘a large literature suggesting that households may not be unitary, and may thus not pool income’ (ibid, p.4). This literature has developed theories of the causal processes of intra-household dynamics involved. The findings of the trial, if internally valid, support Haushofer and Shapiro’s intermediate conclusion that ‘these Kenyan households are more efficient than found in Udry (1996) or Duflo and Udry (2004)’. However, by not developing a causal model of intra-household bargaining in the test population the authors have forfeited the possibility of

contributing to the relevant theoretical literature. A realist RCT could have been embedded in a mixed-methods research programme which used qualitative methods such as focus-grouping to generate hypotheses about difference or sameness of spending patterns between households where transfers were sent to the male head versus those where transfers were sent to the female head. The very modest conclusion of Haushofer and Shapiro that unconditional cash transfers *may* not have perverse effects could already have been supported *a priori*. In order to provide policy makers with evidence about where such perverse effects are less likely, a more realist design would have been necessary.

Despite the problems of external validity discussed above, this trial does make some useful contributions to the wider literature in the form of some results which were counter to the authors' expectations. These are identified as intriguing puzzles and the authors suggest some 'questions for future research'. This motivates the observation that a highly successionist trial with a thin causal model can nevertheless generate surprising results which create knowledge problems. However, it seems unlikely that researchers from outside of development economics will be attracted to these knowledge puzzles given their framing in economics jargon. By contrast, a move to realist RCTs would involve the creation of typologies of CMO combinations, providing 'boundary objects' to facilitate exchange between researchers within different knowledge communities.

5 – CONCLUSION

Chapter two argued that the theoretical basis for the adoption of realist RCTs is strong. Chapter three argued that the practical implementation of theoretical insights is not straightforward and outlined six suggestions for the design of realist RCTs. The critique of the two case studies in chapter four illustrated that applying these suggestions and designing realist RCTs would offer advantages over RCTs underpinned by a successionist account of causation. It is hoped that this critique might introduce some helpful 'anti-discipline' to researchers working within an implicitly successionist knowledge community. However, it is acknowledged that the amount of theorising and modelling this dissertation has argued is necessary for a realist RCT may seem very demanding. For example and as detailed in section three, elaborating a causal model of the test population which meets Cartwright's (2008) requirements is arduous. Unfortunately, there is no real alternative. The promise of theory-free data generation which some see in RCTs is entirely illusory. This is because without a causal model of the test population none of the trial's findings can be rigorously generalised. The policy recommendations of both case studies could only be justified based on some causal model which the trial designers leave largely implicit, forcing policy makers to bet 'willy-nilly' on said causal model (ibid, p.10). Cartwright (ibid) makes the point that this is true 'whether we wish to think about it or not'. Developing explicit causal models is costly but it allows for the generation of typologies of context-mechanism-outcome combinations in which causally-relevant contextual factors are taken explicitly into account. Similarly, being explicit about the underlying justificatory normative framework of an intervention is hard work, as is identifying pathway variables to be tested or testing the components of a complex intervention separately as well as in combination. However, the social world is complex, and it should not surprise us that learning about it is hard work (Lawrence and Després, 2004). As Cartwright (2008, p.41) suggests, '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.'

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 Adato, M. and Hoddinott, J. (Eds.), *Conditional Cash Transfers in Latin America*, Johns Hopkins University Press, pp. 26–54.

Allen, J. and McDowell, L. (1989), *Landlords and property: social relations in the private rented sector*, Cambridge human geography, Cambridge University Press, Cambridge.

Astbury, B. and Leeuw, F.L. (2010), "Unpacking Black Boxes: Mechanisms and Theory Building in Evaluation", *American Journal of Evaluation*, Vol. 31 No. 3, pp. 363–381.

Baird, S.J., Garfein, R.S., McIntosh, C.T. and Özler, B. (2012), "Effect of a cash transfer programme for schooling on prevalence of HIV and herpes simplex type 2 in Malawi: a cluster randomised trial", *The Lancet*, Vol. 379 No. 9823, pp. 1320–1329.

Banerjee, A.V. and Duflo, E. (2008), *The Experimental Approach to Development Economics* (Working Paper No. 14467), National Bureau of Economic Research, available at: http://www.nber.org/papers/w14467 (accessed 31 January 2014).

Basu, K. (2013), *The Method of Randomization and the Role of Reasoned Intuition* (Policy Research Working Paper No. 6722), Washington D.C.: World Bank.

Bhaskar, R. (1975), *A Realist theory of science*, Leeds Books, Leeds.

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*, Vol. 75 No. 12, pp. 2299–2306.

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*, Vol. 94, pp. 81–82.

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*, Vol. 321 No. 7262, pp. 694–696.

Cartwright, N. (2007), "Are RCTs the Gold Standard?", *BioSocieties*, Vol. 2 No. 1, pp. 11–20.

Cartwright, N. (2008), *A Theory of Evidence for Evidence-Based Policy* (Technical Report No. 08/08), Centre for the Philosophy of Natural; and Social Science Contingency and Dissent in Science.

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

Connelly, J. (2004), "Realism in evidence based medicine: interpreting the randomised controlled trial", *Journal of Health Organization and Management*, Vol. 18 No. 2, pp. 70–81.

Craig, P., Dieppe, P., Macintyre, S., Michie, S., Nazareth, I. and Petticrew, M. (2008), *Developing and evaluating complex interventions: new guidance*, Medical Research Council, available at: http://www.mrc.ac.uk/complexinterventionsguidance (accessed 10 September 2014).

Deaton, A. (2010), "Instruments, Randomization, and Learning about Development", *Journal of Economic Literature*, Vol. 48 No. 2, pp. 424–455.

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

Fine, B. (2002), "Economics Imperialism and the New Development Economics as Kuhnian Paradigm Shift?", *World Development*, Vol. 30 No. 12, pp. 2057–2070.

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

Goldstein, J. (2013,August13), "Is It Nuts to Give to the Poor Without Strings Attached?", *The New York Times*, available at: http://www.nytimes.com/2013/08/18/magazine/is-it-nuts-to-give-to-the-poor-without-strings-attached.html (accessed 29 January 2014).

Harré, R. (1970), *The principles of scientific thinking*, Macmillan.

Harré, R. (1985), *The philosophies of science*, Oxford University Press, Oxford; New York, 2nd ed.

Harriss, J. (2002), "The case for cross-disciplinary approaches in international development", *World Development*, Vol. 30 No. 3, pp. 487–496.

Harzing, A.-W. (2010), *Citation analysis across disciplines: The impact of different data sources and citation metrics* (White paper), available at: http://www.harzing.com/data_metrics_comparison.htm (accessed 11 September 2014).

Haushofer, J. and Shapiro, J. (2013a), *Household Response to Income Changes: Evidence from an Unconditional Cash Transfer Program in Kenya* (Unpublished Working Paper).

Haushofer, J. and Shapiro, J. (2013b), *Policy Brief: Impacts of Unconditional Cash Transfers* (Unpublished Working Paper).

Haushofer, J. and Shapiro, J. (2013c), *Welfare Effects of Unconditonal Cash Transfers: Pre-Analysis Plan* (Unpublished Working Paper).

Heckmann, J. (1991), *Randomization and social policy evaluation* (Technical Working Paper No. 107), National Bureau of Economic Research.

Hume, D. (1748), *An Enquiry Concerning Human Understanding*, Oxford University Press.

Karnofsky, H. (2013), “Rigorous study of GiveDirectly’s cash transfers | The GiveWell Blog”, available at: http://blog.givewell.org/2013/12/02/rigorous-study-of-givedirectlys-cash-transfers/ (accessed 29 January 2014).

Kestenbaum, D. (2013), “What Happens When You Just Give Money To Poor People?”, *NPR.org*, available at: http://www.npr.org/blogs/money/2013/10/25/240590433/what-happens-when-you-just-give-money-to-poor-people (accessed 29 January 2014).

Klein, J.T. (1990), *Interdisciplinarity: history, theory, and practice*, Wayne State University Press, Detroit.

Lawrence, R.J. and Després, C. (2004), “Futures of transdisciplinarity”, *Futures*, Vol. 36 No. 4, pp. 397–405.

LéLé, S. and Norgaard, R.B. (2005), “Practicing interdisciplinarity”, *BioScience*, Vol. 55 No. 11, p. 967.

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*, Vol. 339 No. sep10 1, pp. b3496–b3496.

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*, Vol. 94, pp. 124–128.

Mayoux, L. (2006), “Quantitative, Qualitative or Participatory? Which Method, for What and When?”, in Desai, V. and Potter, R.B. (Eds.),*Doing development research*, SAGE, London, pp. 115–129.

Mollinga, P.P. (2010), “Boundary work and the complexity of natural resources management”, *Crop Science*, Vol. 50 No. Supplement 1, p. S–1–S–9.

Oakley, A. (2000), *Experiments in knowing: gender and method in the social sciences*, Polity Press, Cambridge UK.

Pawson, R. and Tilley, N. (1997), *Realistic evaluation*, Sage, London ; Thousand Oaks, Calif.

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”, *Nursing Inquiry*, Vol. 19 No. 1, pp. 18–28.

Rajan, R. and Subramanian, A. (2007), “Does Aid Affect Governance”, *American Economic Review*, Vol. 97 No. 2, pp. 322–327.

Ravallion, M. (2009), “Evaluation in the Practice of Development”, *The World Bank Research Observer*, Vol. 24 No. 1, pp. 29–53.

Saavedra, J. and Garcia, S. (2013), *Educational Impacts and Cost-Effectiveness of Conditional Cash Transfer Programs in Developing Countries: A Meta-Analysis* (SSRN Scholarly Paper No. ID 2333946), Rochester, NY: Social Science Research Network, available at: http://papers.ssrn.com/abstract=2333946 (accessed 31 January 2014).

Sanderson, I. (2000), "Evaluation in Complex Policy Systems", *Evaluation*, Vol. 6 No. 4, pp. 433–454.

Sayer, R.A. (1992), *Method in social science: a realist approach*, Routledge, London; New York, 2nd ed.

Sen, A. (1999), *Development as Freedom*, Oxford University Press, Oxford.

Shaw, I. (1999), *Qualitative Evaluation*, SAGE.

Singal, A.G., Higgins, P.D.R. and Waljee, A.K. (2014), "A Primer on Effectiveness and Efficacy Trials", *Clinical and Translational Gastroenterology*, Vol. 5 No. 1, p. e45.

The Economist (2013), "Pennies from heaven"., *The Economist*, available at: http://www.economist.com/news/international/21588385-giving-money-directly-poor-people-works-surprisingly-well-it-cannot-deal (accessed 29 January 2014).

Westhorp, G. (2014), *Realist impact evaluation: an introduction* (Methods Lab report), Overseas Development Institute, the Australian Department of Foreign Affairs and Trade, BetterEvaluation.

Whatmore, S., Munton, R., Little, J. and Marsden, T. (1987), "Towards a Typology of Farm Businesses in Contemporary British Agriculture Sarah Whatmore Richard Munton Jo Little", *Sociologia Ruralis*, Vol. 27 No. 1, pp. 21–37.

White, H. (2011), *An introduction to the use of randomized controlled trials to evaluate development interventions* (Working Paper No. 9), International Initiative for Impact Evaluation.

Williams, R. (1983), *Keywords: a vocabulary of culture and society*, Fontana Paperbacks, London.

Wong, G., Greenhalgh, T., Westhorp, G., Buckingham, J. and Pawson, R. (2013), "RAMESES publication standards: meta-narrative reviews", *BMC Medicine*, Vol. 11 No. 1, p. 20.

Wong, G., Greenhalgh, T., Westhorp, G. and Pawson, R. (2013), *Realist synthesis: RAMESES training materials* (No. 10/1008/07), National Institute for Health Research Health Services and Delivery Research Program (NIHR HS&DR), p. 54.

World Bank (2009), *Handbook on poverty and inequality*., World Bank, Washington, DC.

Worrall, J. (2007), "Evidence in Medicine and Evidence-Based Medicine", *Philosophy Compass*, Vol. 2 No. 6, pp. 981–1022.

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*, Vol. 21 No. 1, pp. 51–74.