Realist randomised controlled trials: a new approach to evaluating complex public health interventions

Randomized trials of complex public health interventions generally aim to identify what works, accrediting specific intervention 'products' as effective. This approach often fails to give sufficient consideration to how intervention components interact with each other and with local context. 'Realists' argue that trials misunderstand the scientific method, offer only a 'successionist' approach to causation, which brackets out the complexity of social causation, and fail to ask which interventions work *for whom and under what circumstances*. We counter-argue that trials are useful in evaluating social interventions because randomized control groups actually take proper *account of* rather than *bracket out* the complexity of social causation. Nonetheless, realists are right to stress understanding of 'what works for whom and under what circumstances' and to argue for the importance of theorizing and empirically examining underlying mechanisms. We propose that these aims can be (and sometimes already are) examined within randomized trials. Such 'realist' trials should aim to: examine the effects of intervention components separately and in combination, for example using multi-arm studies and factorial trials; explore mechanisms of change, for example analysing how pathway variables mediate intervention effects; use multiple trials across contexts to test how intervention effects vary with context; draw on complementary qualitative and quantitative data; and be oriented towards building and validating 'mid-level' program theories which would set out how interventions interact with context to produce outcomes. This last suggestion resonates with recent suggestions that, in delivering truly 'complex' interventions, fidelity is important not so much in terms of precise activities but, rather, key intervention 'processes' and 'functions'.

Realist trials would additionally determine the validity of program theory rather than only examining ‘what works’ to better inform policy and practice in the long-term.

Research highlights:

1. ‘Realists’ argue that RCTs ignore the complexity of causation and fail to ask which interventions work *for whom and when*.
2. We disagree but believe RCTs RCTs should examine for whom and when interventions work.
3. To this end, RCTs could aim to examine intervention mechanisms of change and how intervention effects vary with context.
4. “Realist” RCTs should examine the validity of intervention theory to better inform policy and practice in the long-term.

Keywords: Public health; complex interventions; evaluation; randomized controlled trials; social experiments; realism; generalizability; social epidemiology.

Realist randomised controlled trials: a new approach to evaluating complex public health interventions

Introduction

In this paper, we outline problems with the way complex public health interventions are sometimes evaluated using randomized controlled trials (RCTs) before examining 'realist' critiques of and proposed alternatives to RCTs. Realism in evaluation represents a paradigm 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 (Kazi, 2003). Realistic evaluators argue that RCTs fail to test hypotheses rooted in theory and embrace a crude notion of causality based on comparison groups and statistical association rather than understanding mechanisms. They argue that evaluators must develop a priori theories about how, for whom and under what conditions interventions will work and then use observational data to examine how context and intervention mechanism interact to generate outcomes. While we dispute the realists' rejection of experimental designs in the social sciences (Pawson & Tilley, 1997), we agree with their arguments concerning the need for evaluation: to examine how, why and for whom interventions work; to give more attention to context; and to focus on the elaboration and validation of program theory. Some previous authors (Blackwood et al., 2010) have argued for a synergistic, rather than oppositional, relationship between realist and randomized evaluation:

The RCT can be used to ascertain whether, all other things being equal, a particular causal mechanism (intervention) is efficacious [i.e. effective under optimum conditions], while realistic evaluation can establish what effect the interaction of other mechanisms operating in the open contexts studied has upon its effectiveness, and identify which mechanisms promote, and which inhibit that effectiveness
(Blackwood et al., 2010, p. 519).

We would go further than this, however, to propose that RCTs themselves could contribute to a realist approach to evaluation. We examine the extent to which some RCTs are already embracing many of these issues and, bringing together some of these existing innovations alongside our own ideas, sketch out what 'realist RCTs' might look like. We argue that it is possible to benefit from the insights provided by realist evaluation without relinquishing the RCT as the best means of examining intervention causality.

Current approaches to evaluating complex public health interventions

RCTs aim to 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. RCTs may randomly allocate individuals or 'clusters' of individuals, such as schools or villages; a method that should ensure that the groups are similar other than differences that occur due to chance. Random allocation is widely regarded as ethical if there is uncertainty about whether intervention confers significant benefits (Bonell et al., 2003). There are formidable challenges to conducting RCTs to evaluate the impacts of complex interventions. Stakeholders, for instance, may resist RCTs out of a belief that randomly assigning participants could unduly deny some the expected benefits of an intervention, even if those benefits have not been demonstrated through rigorous evaluation. The results of RCTs of complex interventions may in some cases be diluted by 'contamination' effects, such as participants assigned to a control group participating in intervention group activities or other services that are similar to the intervention studied. Moreover, whereas RCTs of, for example,

pharmacological interventions are commonly double-blinded (neither provider nor patient is aware to which arm the patient has been allocated), this is rarely the case with social interventions where such blinding is typically impractical. Thus, information bias in RCTs of complex social interventions is more likely. Despite these limitations, we support the view that RCTs provide the strongest evidence about the causal effects of social interventions and are generally feasible except in situations, for example, where intervention delivery is already universal or the pattern decided. They are generally ethical except in situations where some important intervention benefits are already known. In such cases, quasi-experimental designs that form comparison groups based on methods other than randomization, such as 'natural experiments' or trials using statistical matching techniques may be appropriate (Bonell et al., 2011; Craig et al., 2012).

The predominant current approach to trialling complex public health interventions aims to identify which interventions work and replicate those that do via translation studies (Craig et al., 2008). There is concern that RCTs designed primarily to identify whether or not a specific intervention is effective have focussed too much on the internal validity of the trial, addressing the question of efficacy rather than broader questions of reach, effectiveness, adoption, implementation and maintenance (Glasgow et al., 2006). This has led to an evidence base that is dominated by high quality RCTs of poorly theorised interventions, with effects that are poorly understood and unlikely to be universally replicated in translation studies or real world implementation. This model of evidence generation is oriented towards 'accrediting' as effective specific intervention 'products'. This is quite explicit, for example, in the conclusions of some systematic reviews (Farrington & Ttofi, 2010) and the work of organisations such as Blueprints

for Violence Prevention (Center for the Study and Prevention of Violence, 2011). Public health trialists do recognise a balance between maintaining fidelity and enabling adaptation of interventions (Breitenstein et al., 2010; Dane & Schneider, 1998), and some suggest that maintaining the integrity of an intervention's key *functions* (the elements in the process of change that the intervention components aim to facilitate) should be more important than maintaining the integrity of the specific actions used to achieve these (Durlak, 1998; Hawe et al., 2004a). Nonetheless, these debates retain their focus on interventions as potentially generalizable products. The validation of theory does not generally receive the same emphasis in randomized trials or systematic reviews of complex public health interventions, although there are exceptions.

We argue that this product-oriented focus may not be appropriate because complex social interventions are different from other interventions, such as pharmacological ones. First, by definition, complex social interventions combine multiple, synergistic components, which are hypothesized to interact so that the sum of their effects is greater than the effects of their individual parts. One set of guidance suggests that "the greater the difficulty in defining precisely what exactly are the 'active ingredients' of an intervention and how they relate to each other, the greater the likelihood that you are dealing with a complex intervention" (Medical Research Council, 2000). Consider the example of the Intervention with Microfinance for AIDS and Gender Empowerment (IMAGE) intervention, which aims to reduce HIV infections among poor women and their children in rural South Africa by providing HIV health-education workshops, empowerment through peer-led community-development projects, and poverty relief through

microfinance (Pronyk et al., 2006). IMAGE is intended to work via an interaction of these components (Hargreaves et al., 2008).

Second, complex interventions interact with context, meaning that their effectiveness will be dependent upon factors such as socio-economic and environmental conditions, organisational readiness, policy context, target population (Bonell et al., 2006). Thus, the program theory needs to incorporate both the intervention theory and also an understanding of how the intervention interacts with context (Weiss, 1995). This is because local capacity to implement, as well as benefit from, such complex interventions varies and because such interventions usually exert effects via extended causal pathways, which play out differently in different settings.

Consider the example of youth development as a means to reduce teenage pregnancies. Such interventions, comprising mentoring, supplementary education on academic and life skills, and group activities to strengthen self-esteem and aspirations, have been found to be effective in New York City but not all parts of the USA (Kirby et al., 2005; Philliber et al., 2003). They may even have increased rates of teenage pregnancy in England (Wiggins et al., 2009). Such variations in effect might plausibly be explained not only by variations in capacity and fidelity (e.g., programs outside of New York were not as well delivered so may have been less likely to bring benefits), but also by differences in how the interventions interact with mainstream services in different contexts (e.g., in England, the program was often an alternative rather than supplementary to normal schooling services; consequently participation may have caused young people to miss out educationally and feel labelled) and in the social determinants through which both intended and unintended effects operate (e.g., whereas in the USA, the intervention was delivered to all young

people living in areas of dense deprivation, in England, where poverty is less spatially concentrated, the intervention was targeted to individuals judged at risk and this may have brought unintended harms via social-network effects and 'positive deviancy training' (Dishion et al., 1999)).

Currently, RCTs rarely sufficiently recognize these two facts about complex interventions. The first problem is that trials can fail to explore how intervention components and their mechanisms of change interact. Most trialists of complex interventions understandably want to optimize the effectiveness of their intervention, motivated both by a desire to maximize human wellbeing and maximize perceived scientific impact. Funders are similarly motivated by these goals. Each trial team, this paper's authors included, therefore combines an array of intervention components into a single intervention program informed by existing practice, previous evaluation evidence and the team's own values. This tendency has recently been amplified by interventions that are highly branded and sometimes copyrighted as commercial products requiring payment for replication (Botvin et al., 1995; Schaeffer & Bourduin, 2005). The overall result is a multitude of trial teams working without coordination, testing various intervention packages of varying degrees of similarity and difference. Some of these interventions are reported as effective, some as ineffective and a few as harmful, and it is generally difficult if not impossible for either primary RCTs or systematic reviews to make firm conclusions about which intervention components are likely to have the most potential and which combinations of these will produce the greatest effects (Fletcher et al., 2008; Harden et al., 2009).

RCTs have increasingly come to involve embedded process evaluations that collect varying combinations of quantitative and qualitative data (Oakley et al., 2006). These evaluations are often used to examine which components of complex interventions are 'key ingredients'. For example, the IMAGE study employed a process evaluation of both the initial pilot and subsequent local scale-up to explore the relative influence of intervention components on overall outcomes (Hargreaves et al., 2009). The process evaluation concluded that the community-development component might have been the least influential, although this conclusion was based on the limited implementation of this component rather than evidence regarding specific mechanisms of change, and the evaluation could not judge the extent to which better-implemented components did or did not contribute to program effects. To date, process evaluations' ability to shed further light on mechanisms and active ingredients is constrained both by their general tendency to focus on intervention components rather than mechanisms of change and by the inability of those that do aim to examine mechanisms to test rather than merely develop hypotheses about how individual components impact on outcomes.

The second problem is that RCTs generally do not examine how complex interventions interact with the context in which they are implemented and received. The array of teams that are evaluating a diverse range of overlapping interventions work across a range of sites and populations. While some larger trials are sufficiently powered to examine how intervention effects differ by demographic sub-groups (for example, how youth development interventions benefit girls more than boys (Philliber et al., 2002)), existing studies rarely examine how effects differ by geographical or institutional sites. Perhaps this rarity is due to the challenges of

ensuring the even larger samples that this would generally require, or because contextual diversity has simply been neither sought nor measured within study design.

Some trials do collect rich, qualitative data on context to identify facilitators and barriers to intervention implementation and effectiveness but they cannot make very definitive conclusions solely based on these data. For example, a process evaluation of a peer-delivered sex education intervention in England concluded from qualitative data that the wider school context influenced the quality of peer education in terms of the degree of institutional formality framing peer educators' own styles, as well as by imposing constraints on available classroom space and timetables (Strange et al., 2002). In the absence of using quantitative methods to test such qualitatively-driven hypotheses, however, it is hard to establish causal connections between intervention context, processes, and outcomes. Similarly, although some systematic reviews examine how the effects of complex public health interventions vary by demographic sub-group (Main et al., 2008), systematic reviews rarely if ever attempt to use meta-analysis to relate measured characteristics of context to heterogeneity in intervention effects, probably because primary studies rarely report on context consistently (Armstrong et al., 2011).

The implication of this omission is that it is generally impossible for those reading the resultant evidence from experimental studies to develop a clear sense of the *mechanisms of change* through which effects (intended and unintended) occur, which specific intervention components and combinations are likely to be most (and least) effective, and *in what contexts* and with *whom* such effects will potentially be replicable. This is clearly the case, for example, in recent reviews of RCTs of youth development interventions, which have been shown to be simultaneously

effective, ineffective and harmful in different studies (Catalano et al., 2002; Harden et al., 2009; Kirby, 2007). This provides powerful ammunition for 'realist' critiques of RCTs, as we shall see below.

The 'realist' critique of, and alternative to, RCTs

The realist tradition provides a critique of, and apparent alternative to, the RCT tradition, but we argue it might be more useful in informing how RCTs could be modified to be more useful in the evaluation of complex public health interventions. Realism asserts that the objects of knowledge exist independently of our minds and that unobservable objects and structures can exert causal influence (Bhaskar, 1975), in contrast to a positivist tradition, which would limit science to directly observable phenomena. The most prominent exponents of realist evaluation are criminologists Ray Pawson and Nick Tilley, who criticize the RCT approach for its positivistic assumptions, and suggest alternatives based on a realist perspective. We briefly summarize these criticisms and suggestions before critically evaluating these and proposing a synthesis involving realist RCTs of complex public health interventions.

Pawson and Tilley's first criticism of the experimental approach to evaluation is that it embodies a misunderstanding of the scientific method (Pawson & Tilley, 1997). This misunderstanding is said to lie in the positivistic assumption that empirical research on observable phenomena that is uninformed by prior theories about causal mechanisms can produce scientific knowledge. They argue that, instead, science must: (a) assume the presence of underlying mechanisms of causality, even if these might be unobservable; (b) develop theory about how these operate; and (c) use this

theory to direct empirical work that explores whether the theory can explain observable phenomena. Huygens' work on the physics of pendular motion is held up as an example of good science because, rather than trying to develop theory on the basis of empirical observation, Huygens first developed a theory and then conducted controlled observations to test whether the theory was predictive.

Pawson and Tilley's second criticism of the experimental method is that it embodies a "successionist" notion of causality, whereby cause and effect is suggested by the "repeated succession of one ... event by another" (Pawson & Tilley, 1997, p.5). Regarding the RCT approach, they contend that "(i)n pursuing causal explanation via a constant conjunction model, with its stress on that which can be observed and controlled, it has tended to overlook the liabilities, powers and potentialities of the programs and subjects whose behaviour it seeks to explain" (Pawson & Tilley, 1997, p. 34). In other words, RCTs are too controlled and too disconnected from the interactions of participants and environments, due to the positivistic epistemology that underlies them. This has been described as the 'black box' problem in RCTs of health promotion interventions (Green and Tones, 1999). In contrast, Pawson and Tilley assert that their method of realist evaluation understands causality in "generativist" terms: an elucidation of the mechanism and properties of phenomena that underlie causality. Related to this, they argue that sound scientific enquiry in the natural as well as the social sciences does not employ counterfactuals because of this orientation to examining causality in terms of the internal processes of causality, again citing as an example Huygen's work, which lacked any external control group.

Finally, Pawson and Tilley argue that, in terms of their policy utility, RCTs are misguided in aiming to assess *whether* interventions work because they miss the point that most interventions will work for some people under some conditions, and what evaluators should tell policy-makers is who are these people and what are these conditions. They cite the example of trials of mandatory arrest for domestic violence as a means of reducing rates of repeated assault (Sherman, 1992). While an initial study in Minneapolis reported a statistically significantly lower rate of repeat calls for domestic violence associated with those randomized to be arrested, subsequent studies of mandatory-arrest programs reported mixed findings with some even showing increased risk of repeat domestic violence among the intervention group. Pawson and Tilley suggested that this variation in outcomes can be explained by varying community, employment and family structures in the different cities; for example, arrest is a greater source of social shame and a more powerful stimulus to change in contexts with more enduring pro-social norms (Pawson and Tilley, 1997).

This brings us to Pawson and Tilley's suggested realist mode of evaluation as an alternative to the RCT approach. Following on from their focus on "what works for whom in what circumstances?", they suggest the importance of evaluators understanding and developing theories of underlying causal mechanisms and, in particular, examining how 'context' interacts with intervention 'mechanisms' to generate 'outcomes', to be tested through naturally-occurring, observational data. For example, in evaluating the effectiveness of closed-circuit television (CCTV) in car parks as a means of reducing car crime (Tilley, 1993), they advocate developing multiple theories regarding the mechanisms by which this might work and testing these. For example, car crime could be reduced via: CCTV enabling the arrest, removal and future

deterrence of offenders; encouraging increased use of car parks and greater 'natural surveillance'; or CCTV and its associated signage symbolizing efforts to take crime seriously leading to deterrence of current and potential offenders. They stress that these theories must also address how context might interact with intervention mechanism to determine outcomes. With CCTV, for instance, in some sites high rates of car crime result from a few, very active offenders, so the arrest and removal mechanism could be highly effective in this context. This approach would plausibly be less effective where crime arises from many, occasional offenders. Additionally, in car parks where natural surveillance is already high, CCTV may have less impact. Pawson and Tilley recommend a series of studies drawing on naturally occurring (not experimentally manipulated) data from a range of sites to assess the validity of such underlying context-mechanism-outcome theories, such as measuring conviction rates to test the first theory and temporal patterns of crime across different car parks to test the second hypothesis.

Problems with the 'realist' position

One might question Pawson and Tilley's claim that the use of RCTs inevitably embodies a set of strong epistemological and ontological commitments, such as the positivistic account of scientific knowledge and the 'successionist' theory of causality (Bonell et al., 2003). However, we will not address this broader point here. Instead, firstly, we will concentrate on Pawson and Tilley's critique of the use of counterfactuals to examine causality. Pawson and Tilley are incorrect in arguing that sound experiments in natural science never employ counterfactuals. RCTs are used, for example, in agricultural science (Oakley, 2000). More importantly, they fail

to understand (or at least convey) why experiments in social science do require external counterfactuals and those in some natural sciences do not.

Use of a counterfactual is necessary within experiments involving social interventions on human populations precisely because these sorts of experiments involve much more complex causality where it is not possible to isolate the effects of one causal factor from another. Whereas in physics one can reasonably hypothesize that gravity causes certain forms of motion and can devise experiments to isolate the effects of gravity from other forces, it would not be practical, ethical or useful to do so in social experiments. Instead, one can only hypothesize that a social intervention will *interact* with an array of contextual factors to cause an outcome. Additionally, the magnitude of effects in complex public health interventions can be much more modest, gradual, and entangled with other factors than in the case of clinical treatments or physics experiments that often result in immediate and highly observable responses that can be more obviously attributed to an intervention in the absence of a control group. It is therefore only legitimate to examine this interaction of causal agents by comparing two sets of conditions, both of which containing the aforementioned contextual conditions but only one containing the intervention.

Pawson and Tilley misunderstand this use of control groups, critiquing it as reflecting trialists' belief that they can strip away the context and isolate an intervention's causal effects by *bracketing out* all other contextual factors and mechanisms that might contribute to producing that outcome. Instead, RCTs' use of control groups actually reflects the opposite: how interventions *interact* with contextual factors in order to bring about an outcome. Thus, we argue

that control groups actually take proper account of the influence of other factors rather than trying to obviate them. We also posit that Pawson and Tilley are wrong to suggest that it is not important to ask questions about whether interventions work at all. While it is clear from our introduction that we agree that RCTs have often failed to provide the evidence about how and in what contexts interventions are effective, this does not detract from the importance of RCTs asking the more basic question of whether the introduction of an intervention to a particular population in a particular site produces improvement in outcomes when compared to similar populations and sites without the intervention. Trials have provided evidence of intervention benefits (Berrueta-Clement et al., 1984) as well as harms (Dishion et al., 1999) that might not have been established by other designs with weaker control of confounding factors.

By neglecting the counterfactual, and therefore failing to test hypotheses about what might have happened in the absence of intervention, the model of 'realist' evaluation proposed by Pawson and Tilley is extremely limited. At best, it can only develop a sense of the *plausibility* of the effects of an intervention, not their probability. Consider the case of a non-experimental realist evaluation of the effects of a multi-component health-education, community-development and microfinance intervention aiming to prevent HIV in sub-Saharan South Africa, such as the IMAGE project described above. It would be possible to undertake observational studies that may find that such an intervention is associated with reduced rates of HIV in villages where new infections are concentrated among women under age 20 and where rates of school drop-out were high (suggesting that reducing school drop-out via poverty prevention may be a key causal mechanism). These findings would suggest a very plausible account of how context and intervention mechanism interacted to produce beneficial outcomes. Critically, however, in the

absence of an experimental counterfactual, these accounts could be entirely wrong. For example, it is equally plausible that in the villages where HIV was reduced this arose from confounding factors (perhaps those villages were selected for intervention because they were the communities most determined to reduce rates of HIV), regression to the mean (perhaps the intervention took place in those villages because rates of HIV were extremely high, and HIV rates fell to a more normal level over time quite independently of the intervention), or simply random error.

It is instructive that in most of Pawson and Tilley's examples (real and hypothetical) they actually do involve (non-random) external control groups, including the comparison of car crimes in car parks where CCTV is accompanied by high and low levels of signage. Sometimes, Pawson and Tilley fail to recognize this, such as in their discussion of an evaluation of prison education for reoffending rates where the analysis draws on an unacknowledged external comparison group made up of those prisoners who previously provided the data from which 'predicted reoffending rates' were created for the evaluation participants.

Despite these criticisms of Pawson and Tilley's overall position, we believe they are right to stress that the question for impact evaluation is 'what works for whom and under what circumstances' as well as to argue for an emphasis on theorizing and empirically examining underlying mechanisms and their interaction with context. Nonetheless, we believe these questions can be (and sometimes already are) examined using RCTs. In fact, such evaluation designs provide the most rigorous means of examining questions of mechanisms and context while still enabling a rigorous assessment of causal attribution, which Pawson and Tilley's non-experimental approach lacks.

A synthesis: realist RCTs of complex public health interventions

What might a 'realist' program of RCTs look like? We set out some suggestions below illustrated by cases in which RCTs are already employing these approaches, and we suggest how these may be further enhanced.

First, realist RCTs would place emphasis on understanding the effects of intervention components separately as well as in combination. For example, this could involve use of 'multi-arm studies' with various combinations of intervention components in each arm. Factorial trials (Montgomery et al., 2003) could also be used. Rather than there being one intervention and one comparison group, there are instead two intervention components and four groups: two groups each receiving each of the individual intervention components, one group receiving both interventions and one group receiving neither. Some evaluations already have used such designs (Dangour et al., 2007; Flay et al., 2004), for example to examine the relative effects of school-based health promotion components that facilitate changes to the overall school environment versus curriculum components (Flay et al., 2004). A few systematic reviews have also examined intervention components. For instance, the review of bullying interventions cited above (Farrington & Ttofi, 2010) examined the extent to which specific components, such as disciplinary methods, parent training and cooperative group-work, were associated with greater effect estimates. It might be useful for evaluations focusing on particular components to examine the effects of the single component on the proximal outcomes which that specific component is intended to affect rather than the end-outcomes intended for the overall multi-component

program. Drawing on earlier work by Box (1958) on ‘mini RCTs’ for continuous evolutionary evaluation, Eccles and Templeton (2001) have suggested, for example, that RCTs of youth-development interventions ‘step back’ and evaluate how different approaches to management might affect measures of the quality of delivery rather than the overall intervention’s impact on distal health or social outcomes.

Second, realist RCTs would place emphasis on examining mechanisms of change. For example, they could investigate mechanisms’ effects on immediate theorised benefits, such as improved knowledge, changed attitude and altered behavior, rather than on the ultimate, longer-term health outcomes that an intervention aspires to change. This is already a feature of many trial analyses that examine intermediates both as secondary outcomes and also as pathway variables included in models to assess whether they account for intervention effects on primary outcomes. One school-based youth development program to reduce young people’s smoking measured increased self-efficacy as a secondary outcome and potential pathway to tobacco-avoiding behavior (Winkleby et al., 2004). Methods have been established for some time regarding how to assess mediation statistically (Baron & Kenny, 1986; MacKinon et al., 2002). Mediation analysis may also shed light on which intervention components are ‘active ingredients’ within simple two-arm RCTs. Gardner and colleagues, for example, reported that parenting skills, rather than parental confidence, appear to mediate the effects of “Incredible Years” parenting classes on child behavioural outcomes, suggesting that the intervention component responsible for the intervention’s overall effects is the provision of skills rather than the component by which intervention recipients provide peer support to one another (Gardner et al., 2006).

Third, we propose that there could be a more strategic, coordinated approach to testing the effects of interventions and their components in different contexts using consistent measures where possible (Breitenstein et al., 2010). Although this purposive contrasting by context does occur *within* some studies, for example, within cluster RCTs involving a diversity of clusters and populations in order to examine sub-group effects (Strange et al., 2002), we are unaware of it happening in any pre-planned manner between studies. Some reviews have attempted to examine how intervention effects vary by contextual factors, but these are undermined by the poor reporting of intervention content and theory (Armstrong et al., 2011), and by the lack of consistent measures for describing contexts. For example, while Lipsey et al examined how the effectiveness of youth offender rehabilitation interventions varied by the baseline risk of participants as well as allocation to different arms of the judicial service (because these aspects of context are relatively consistently described), they were unable to examine heterogeneity of effect by other, less-well-described aspects of context, such as the presence of local product champions or well-resourced collaborating institutions (Lipsey, 2009). While more consistent measurement of context would be useful, we recommend that greater theorization and pre-development of hypotheses regarding what aspects of context are likely to be important for what sorts of interventions so that studies can be planned prospectively to examine these possibilities. The hypotheses to be tested would take the form of the context-mechanism-outcome configurations suggested by Pawson and Tilley. This would require coordinated programs of evaluations oriented toward the testing of common theories.

Fourth, realist RCTs should involve interchange between qualitative and quantitative research methods (Palinkas et al., 2010). Qualitative research can develop hypotheses about which

intervention components are most important, how the intervention works and how context affects implementation and impacts. For example, qualitative research, drawing on observations, interviews or focus groups, is essential to identify: variation in delivery of an intervention (Fagan et al., 2008; Hill et al., 2007); context helping or hindering delivery (Hawe et al., 2004b) or effect (Moore & Tapper, 2008); or participant interaction with the intervention (Heaven et al., 2006). But longitudinal quantitative data are then also required to test the wider relevance of findings from qualitative research. For example, a realist RCT might collect qualitative data in the period between baseline and follow-up questionnaire surveys and use emergent findings from the qualitative research to inform additional quantitative measures to be included in follow-up surveys.

Our fifth recommendation is that a crucial aim of realist RCTs would not only be the 'accreditation' of specific intervention 'products' as effective but also the building and validation of program theories of interventions. These would be mid-level theories that, as Pawson and Tilley suggest, would aim to explain how context and an intervention's underlying mechanisms interact to produce outcomes. They would build on 'logic models' that define the components and mechanisms of specific interventions within a very particular setting (referred to as implementation models by Weiss, 1995), but they also would provide more consideration of how intervention mechanisms interact with context (Connell & Kubisch, 1998; Patton, 2002; Weiss, 1995). As such, it is essential that interventions be developed with a clearly articulated theory of change (Connell & Kubisch, 1998). Investigators would describe the theory of change in evaluation reporting as well as exactly how the given study aims to examine the theory of change. By including contextual factors within these theories, they would explicitly consider the

circumstances in which the intervention may be optimized or may not work at all, and thus be invaluable in defining optimal targeting, tailoring, implementation and organizational support.

Consider the following example. A cluster RCT of peer-education in schools to prevent smoking has suggested the effectiveness of the 'ASSIST' intervention (Campbell et al., 2008). In the short term, the main value to public health and education policy-makers is likely to be in the accreditation of the ASSIST 'product' as an effective response to youth smoking (National Institute for Health and Clinical Excellence, 2010). This conclusion might not, however, be the most important one for policy-makers over the longer term because the context in which ASSIST operated will inevitably start to alter. Schools and the education system in which they operate will evolve, affecting their capacity to deliver the intervention and the relevance of some of the specific components of the ASSIST package to the new context. The culture of students and how they communicate will also evolve, potentially affecting the acceptability of the specific intervention. Most importantly of all (and partly as a result of the very delivery of the ASSIST intervention), the precise sets of needs that underlie young people's vulnerability to smoking may change. For instance, peer norms might become much less influential compared to other determinants. As a result, the specific intervention product accredited as effective, in this case ASSIST, might wane in its feasibility and acceptability and become less effective, and the specific form of the various intervention components may need to be updated to successfully deliver the key functions of the main theoretical components. A history of what works in public health is precisely that: a history of success, not a guarantee for the future.

In the longer term, however, it might be that the main value of the ASSIST trial is its contribution to intervention theory, particularly concerning the process of identifying peer opinion leaders and diffusing health promoting messages and norms through school based social networks to change behaviour (Starkey et al., 2009). More broadly, the ASSIST findings support the theory of the 'diffusion of innovations' (Rogers, 1962), and further trials are needed to understand in what contexts the peer-led approach, embodied by ASSIST, engages with population needs in order to bring about beneficial effects, as well as which specific components appear to work best to facilitate this overall approach in different settings and whether this model can be used to target other behaviours.

Nonetheless, we would also like to stress that realist RCTs should still aim to generate evidence about what particular interventions achieve what particular outcomes. Such evidence would be extremely useful in informing replication in similar settings, particularly in the short-to-medium-term. What we argue, however, is that investigators employing realist RCTs should accept the limited transferability across time and space of particular methods and, as such, attempt also to examine how well particular intervention theories play out in different contexts in order to provide more useful information about generalizability across broader leaps of time and space.

We believe it is more useful to policy-makers and practitioners for evaluators additionally to assess the validity of these mid-level program theories, rather than only to determine which specific interventions are effective, because of the more uncertain generalizability of the latter (Bonell et al., 2006). Program implementers might also be more likely to participate in and benefit from RCTs if the evaluation is conducted in such a way that treats them as co-learners for

the development of interventions and program theory, rather than simply threatening their work's existence if a trial does not demonstrate certain outcomes. This is consistent with our experience of working with schools, non-governmental organizations and government (Bonell et al., 2010; Morton & Montgomery, 2012; Murphy et al., 2011).

It would be up to those using these program theories to develop the intervention components, to involve stakeholders in moving from theory to intervention and pilot these interventions to establish if they are locally feasible and acceptable. This resonates with recent suggestions that in delivering truly complex social interventions it is important to maintain fidelity, not so much in terms of precisely what form the intervention is delivered in at each site, but in what intervention 'processes' and 'functions' are initiated in each site (Hawe et al., 2004a). Our aim here is to advance this argument by suggesting that these 'functions' should be defined in terms of program theories that spell out the mechanisms of change of intervention components.

Conclusions

This paper has scrutinized apparent tensions between RCT and realist approaches to evaluating complex public health interventions and proposed opportunities for synergy between the two. Some argue against anything other than an approach to complex interventions being experimentally evaluated and those found to be effective being replicated with maximum fidelity because we have so little understanding of key causal mechanisms (Kemple & Willner, 2008). However, given the uncertain and mixed evidence for whether fidelity or adaptation in replication is most likely to promote effectiveness (Barber et al., 2006; Blakely et al., 1987; Dane

and Schneider, 1998; Durlak & Dupre, 2008; McLaughlin, 2000) even in the case of less complex interventions, and given the disappointing results of many replication studies of complex interventions even where fidelity was well-maintained, such as the CAS Carrera project cited above, we recommend that trialists recognise the importance of understanding causal mechanism theories and the extent to which these can be generalized between contexts.

We conclude that, while there are undoubted challenges with conventional approaches to experimental evaluation of complex public health interventions, there are even more serious ones with realist approaches to evaluations that reject experimental methods altogether. The philosophy of realism in the social sciences should not rule out use of experimental methods. RCTs are crucial both to answer cause-effect questions about public health and to build and refine theory.

It might be argued that our plans are impossible because of the expense and time required for such painstaking assessment of how different intervention components play out in different sites. In many cases, institutions funding impact evaluations may not supply the level of resources required for more intensive, realist RCTs. However, we argue that this systematic approach would ultimately be more time- and cost-efficient than current, uncoordinated efforts. There are already huge numbers of evaluation studies conducted that focus on an unsystematic array of combinations of intervention components and settings, and which, for the reasons outlined above, often provide a poor basis for decision-making by policy-makers and practitioners (Kessler & Glasgow, 2011). Some might prefer a more mechanical approach to conducting and interpreting RCTs of packaged interventions that largely or fully neglects careful attention to

theory. Yet, though simpler, this approach does not draw out the best of what RCTs have to offer researchers, decision-makers, and program-implementing staff.

References

Armstrong, R., E. Waters, et al. (2008). Improving the reporting of public health intervention research: advancing TREND and CONSORT. *Journal of Public Health*, 30(1), 103-9.Barber, J. P., R. Gallop, et al. (2006). The role of therapist adherence, therapist competence and allowance in predicting outcome of individual drug counselling: results from the national institute drug abuse collaborative cocaine treatment study. *Psychotherapy Research*, 16, 229-240.Baron, R. M. and D. A. Kenny (1986). The moderator-mediator variable distinction in social psychological research: conceptual, strategic and statistical considerations. *Journal of Personality and Social Psychology*, 51, 1173-1182.

Berrueta-Clement, J., L. Schweinhart, et al. (1984). *Changed Lives: The effects of the Perry Preschool Program on youths through age 19.* Ypsilanti: The High/Scope Press.

Bhaskar, R.A. (1975). *A Realist Theory of Science.* London: Verso.

Blackwood, B., P. O'Halloran, et al. (2010). On the problems of mixing RCTs with qualitative research: the case of the MRC framework for the evaluation of complex healthcare interventions. *Journal of Research in Nursing*, 15(6), 511-521.

Blakely, C., J. P. Mayer, et al. (1987). The fidelity-adaptation debate: implications for the implementation of public sector social programs. *American Community Psychology*, 15, 253-268.

Bonell, C., Bennett, R., Oakley, A. (2003). Sexual health interventions should be subject to experimental evaluation. In J. Stephenson, J. Imrie and C. Bonell (Eds.), *Effective Sexual Health Interventions: Issues in Experimental Evaluation*. Oxford: Oxford University Press.

Bonell, C., A. Oakley, et al. (2006). Trials of health interventions and empirical assessment of generalizability: suggested framework and systematic review. *British Medical Journal*, 333, 346-349.

Bonell, C., A. Sorhaindo, et al. (2010). Pilot multi-method trial of a school-ethos intervention to reduce substance use: building hypotheses about upstream pathways to prevention. *Journal of Adolescent Health,* 47(6), 555-563.

Bonell C, Hargreaves J, et al. (2011). Alternatives to randomisation in the evaluation of public-health interventions: design challenges and solutions. *Journal of Epidemiology and Community Health,* 65(7), 582-7.

Botvin, G. J., E. Baker, et al. (1995). Long-term Follow-up Results of a Randomized Drug Abuse Prevention Trial in a White Middle-class Population. *Journal of the American Medical Association*, 273(14), 1106-1112.

Box, G. E. P. (1958). Evolutionary Operation: A Method for Increasing Industrial Productivity. *Applied Statistics*, 6, 81-101.

Breitenstein, S. M., D. Gross, et al. (2010). Implementation fidelity in community-based Interventions. *Research in Nursing & Health*, 33, 164-173.

Campbell, R., F. Starkey, et al. (2008). An informal school-based peer-led intervention for smoking prevention in adolescence (ASSIST): a cluster randomised trial. *The Lancet*, 371, 1595-1602.

Catalano, R. F., L. M. Berglund, et al. (2002). Positive Youth Development in the United States: Research Findings on Evaluations of Positive Youth Development Programs. *Prevention & Treatment*, 5(1), 1-166.

Center for the Study and Prevention of Violence (2011). Retrieved 20 December, 2011, from http://www.colorado.edu/cspv/blueprints/.

Connell, J. P. and A. C. Kubisch (1998). *Applying a Theory of Change Approach to the Evaluation of Comprehensive Community Initiatives: Progress, Prospects, and Problems.* Washington DC: The Aspen Institute.

Craig, P., P. Dieppe, et al. (2008). Developing and evaluating complex interventions: the new Medical Research Council guidance. *British Medical Journal*, 337, a1655.

Craig, P., C. Cooper, et al. (2012). Using natural experiments to evaluate population health interventions: new Medical Research Council guidance. *Journal of Epidemiology and Community Health*, May 10. [Epub ahead of print, doi:10.1136/jech-2011-200375]

Dane, U. A. and B. H. Schneider (1998). Program integrity in primary and early secondary prevention: are implementation effects out of control? *Clinical Psychology Review*, 18(1), 23-45.

Dangour, A., C. Albala, et al. (2007). A factorial-design cluster randomised controlled trial investigating the cost-effectiveness of a nutrition supplement and an exercise programme on pneumonia incidence, walking capacity and body mass index in older people living in Santiago, Chile: the CENEX study protocol. *Nutrition Journal*, 6: 14.

Dishion, T. J., J. McCord, et al. (1999). When interventions harm. *American Psychologist*, 54(9), 755-764.

Durlack, J. A. (1998). Why intervention implementation is important. *Journal of Prevention and Intervention in the Community*, 17(2), 5-18.

Durlak, J. A. and E. P. Dupre (2008). Implementation matters: a review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology*, 41, 327-350.

Eccles, J. S. and J. Templeton (2001). Community Based Programs for Youth: Lessons Learned from General Development Research and From Experimental and Quasi-experimental evaluations. *Urban Health Initiative Seminar Series*. Cambridge, Harvard University.

Fagan, A., K. Hanson, et al. (2008). Bridging Science to Practice: Achieving Prevention Program Implementation Fidelity in the Community Youth Development Study. *American Journal of Community Psychology*, 41(3), 235-249.

Farrington, D. P. and M. M. Ttofi (2010). *School-based programs to reduce bullying and victimization, Campbell Systematic Reviews*, 6, 1–147

Flay, B. R., S. Graumlich, et al. (2004). Effects of 2 prevention programs on high-risk behaviors among African American youth: a randomized trial. *Archives of Pediatric and Adolescent Medicine*, 158(4), 377-384.

Fletcher, A., C. Bonell, et al. (2008). School effects on young people's drug use: a systematic review of intervention and observational studies. *Journal of Adolescent Health*, 42(3), 209-220.

Gardner, F., J. Burton, et al. (2006). Randomised controlled trial of a parenting intervention in the voluntary sector for reducing child conduct problems: outcomes and mechanisms of change. *Journal of Child Psychology and Psychiatry*, 47(11), 1123-1132.

Glasgow, R.E., L.M. Klesges, et al. (2006). Evaluating the Overall Impact of Health Promotion Programs: Using the RE-AIM Framework for Decision Making and to Consider Complex Issues. *Health Education Research*, 21(3), 688-694.

Green, J. and K. Tones (1999). Towards a secure evidence base for health promotion. *Journal of Public Health Medicine*, 21, 133-139.

Harden, A., G. Brunton, et al. (2009). Teenage pregnancy and social disadvantage: a systematic review integrating trials and qualitative studies. *British Medical Journal*, 339, b4254.

Hargreaves, J., A. Hatcher, et al. (2009). Process evaluation of the Intervention with Microfinance for AIDS and Gender Equity (IMAGE) in rural South Africa. *Health Education Research*, 25(1), 27-40.

Hargreaves, J. R., C. P. Bonell, et al. (2008). Systematic review exploring time-trends in the association between educational attainment and risk of HIV infection in sub-Saharan Africa. *AIDS*, 22(3), 403-414.

Hawe, P., A. Shiell, et al. (2004a). Complex interventions: how "out of control" can a randomised controlled trial be? *British Medical Journal*, 328, 1561-1563.

Hawe, P., A. Shiell, et al. (2004b). Methods for exploring implementation variation and local context within a cluster randomised community intervention trial. *Journal of Epidemiology and Community Health*, 58(9), 788-793.

Heaven, B., M. Murtagh, et al. (2006). Patients or research subjects? A qualitative study of participation in a randomised controlled trial of a complex intervention. *Patient Education and Counseling*, 62(2), 260-270.

Hill, L., K. Maucione, et al. (2007). A Focused Approach to Assessing Program Fidelity. *Prevention Science*, 8(1), 25-34.

Kazi, M. (2003). Realist Evaluation for Practice. *British Journal of Social Work, 33*(6), 803-818.

Kemple, J. and C. J. Willner (2008). *Career Academies: Long-term impacts on labor market outcomes, educational attainment, and transitions to adulthoood*. New York: MDRC.

Kessler, R., and R.E. Glasgow. (2011). A proposal to speed translation of healthcare research into practice: dramatic change is needed. *American Journal of Preventive Medicine,* 40(6), 637-44.

Kirby, D. (2007). Emerging Answers 2007: *Research Findings on Programs to Reduce Teen Pregnancy and Sexually Transmitted Diseases*. Washington DC: The National Campaign to Prevent Teen and Unplanned Pregnancy.

Kirby, D. B., T. Rhodes, et al. (2005). *Implementation of multi-component youth programs to prevent teen pregnancy modelled after the Children's AID Society - Carrera Program*. Scotts Valley: ETR Associates.

Lipsey, M. W. (2009). The primary factors that characterize effective interventions with juvenile offenders: A meta-analytic overview. *Victims and Offenders*, 4, 124-147.

MacKinon, D. P., C. M. Lockwood, et al. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, 7, 83-104.

Main, C., S. Thomas, et al. (2008). Population tobacco control interventions and their effects on socialinequalities in smoking: placing an equity lens on existing systematic reviews. *BMC Public Health*, 8, 178.

McLaughlin, M. (2000). *Community Counts: How Youth Organizations Matter for Youth*. Washington DC: Public Education Network.

Montgomery, A. A., T. J. Peters, et al. (2003). Design, analysis and presentation of factorial randomised controlled trials. *BMC Medical Research Methodology*, 3, 26.

Moore, L., and K. Tapper K. (2008). The impact of school fruit tuck shops and school food policies on children's fruit consumption: a cluster randomised trial of schools in deprived areas. *Journal of Epidemiology and Community Health*, 62(10), 926-931.

Morton, M. H. and P. Montgomery (2012). Empowerment-based non-formal education for Arab youth: A pilot randomized trial. *Children and Youth Services Review*, 34(2), 417-425.

Medical Research Council (2000). *A framework for the development and evaluation of randomised controlled trials for complex interventions to improve health*. London: Medical Research Council.

Murphy S., G.F. Moore G.F, et al. (2011). Free healthy breakfasts in primary schools: a cluster randomised controlled trial of a policy intervention in Wales. *Public Health Nutrition*, 19(4), 362-374.

National Institute for Health and Clinical Excellence (2010). *School-based interventions to prevent smoking (Quick reference guide)*. London: NICE.

Oakley, A. (2000). *Experiments in Knowing: Gender and Method in the Social Sciences*. Cambridge: Polity.

Oakley, A., V. Strange, et al. (2006). Integrating process evaluation in the design of randomised controlled trials of complex interventions: the example of the RIPPLE Study. *British Medical Journal*, 332, 413-416.

Palinkas, L. A., G. Aarons, et al. (2010). Mixed method designs in implementation research. *Administration and Policy in Mental Health*, 38, 44-53.

Patton, M. Q. (2002). *Qualitative Research & Evaluation Methods* (3 ed.). Thousand Oaks: Sage.

Pawson, R. and N. Tilley (1997). *Realistic Evaluation*. London: Sage.

Philliber, S., J. W. Kaye, et al. (2002). Preventing pregnancy and improving health care access among teenagers: an evaluation of the Children's Aid Society-Carrera Program. *Perspectives on Sexual and Reproductive Health*, 34(5), 244-251.

Pronyk, P. M., J. R. Hargreaves, et al. (2006). Effect of a structural intervention for the prevention of intimate-partner violence and HIV in rural South Africa: a cluster randomised trial. *The Lancet*, 368(9551), 1973-1983.

Rogers, E. M. (1962). *Diffusion of Innovations*. New York: The Free Press.

Schaeffer, C. M. and C. M. Borduin (2005). Long-Term Follow-Up to a Randomized Clinical Trial of Multisystemic Therapy With Serious and Violent Juvenile Offenders. *Journal of Consulting and Clinical Psychology*, 73(3), 445-453.

Sherman, L. (1992). *Policing Domestic Violence*. New York: Free Press.

Starkey, F., S. Audrey, et al. (2009). Identifying influential young people to undertake effective peer-led health promotion: the example of A Stop Smoking In Schools Trial (ASSIST). *Health Education Research*, 24(6), 977-988.

Strange, V., S. Forrest, et al. (2002). What influences peer-led sex education in the classroom? A view from the peer educators. *Health Education Research*, 17(3), 339-349.

Tilley, N. (1993). *Understanding Car Parks, Crime and CCTV: Evaluation Lessons from Safer Cities, Crime Prevention Unit Series Paper 42*. London: Home Office.

Weiss, C. H. (1995). Nothing as Practical as Good Theory: Exploring Theory-based Evaluation for Comprehensive Community Initiatives for Children and Families. In J. P. Connell, A. C. Kubisch, S. L.B. and C. H. Weiss (Eds.), *New Approaches to Evaluating Community Initiatives: Concepts, Methods, and Contexts*. Washington DC: Aspen Institute.

Wiggins, M., C. Bonell, et al. (2009). Health outcomes of youth development programme in England: prospective matched comparison study. *British Medical Journal*, 339, b2534.

Winkleby, M. A., E. Feighery, et al. (2004). Effects of an advocacy intervention to reduce smoking among teenagers. *Archives of Pediatric and Adolescent Medicine*, 158(3), 269-275.