The case for experimental design in realist evaluation

Andrew Hawkins
ARTD Consultants
andrew.hawkins@artd.com.au

Keywords: realist evaluation, experimental evaluation, realist experimental design, transfactual mechanisms, demi-regularities, effect size, testing CMO (Context-Mechanism-Outcome) configurations, realist cost benefit analysis, youth mentoring

Suggested citation: Hawkins, A.J. (2014). The case for experimental design in realist evaluation. *Learning Communities: International Journal of Learning in Social Contexts, Special Issue: Evaluation, 14, 46-59.*

Abstract

This article argues the realist critique of experimental design to evaluate interventions in complex social systems is valid but incomplete. It argues for experimental approaches to testing realist theory and for estimating effect sizes.

The paper aims to provide a means for scientific understanding of the relative value of interventions in different contexts, for whom and to what effect. The paper is grounded in a realist philosophy of science and a realist approach to evaluation. It argues for the use of experimental design to test and estimate the magnitude of an outcome in a hypothesised realist Context-Mechanism-Outcome (CMO) configuration. The approach requires that program theory (rather than the program) is the unit of analysis. It also requires that context – crucial for a mechanism firing – is brought into the effect size equation, while at the same time attempts are made to control for the effects of other mechanisms.

The focus of this paper is on the general approach rather than a particular method. The approach was applied in an evaluation of a youth mentoring program. The method used was a matched-pair, pre and post-test, control group quasi-experimental design. The results of our application of the approach were limited but provided insight about the extent to which a particular mentoring mechanism, when properly targeted, could generate outcomes for certain students.

This approach to evaluation is consistent with underlying principles of scientific realism and theory testing and provides a means for generating evidence about the value of interventions in complex social systems, for whom and to what extent.

Introduction

Evaluators and policy makers are both concerned with understanding how to design and target interventions for maximum effect and understanding the relative worth of interventions. In many fields of research, "the controlled experiment is king" (Jeffery-Evans, 2012, p. 26) and the randomised control trial (RCT) is considered the gold standard, the top of the evidence hierarchy or the "the best way of determining whether a policy is working" (Haynes, 2012, p. 4). The concern here is with the relative worth of whole programs, about answering the question 'does this program work?' and 'to what extent?' While there is growing interest in the concept of mechanisms across many areas of social science (Astubry & Leeuw, 2010, p. 363), experimental evaluators tend to better meet the demands of policy makers for summative evaluations and cost benefit analyses of particular programs.

The realist does not ask 'what works?', but 'what works for whom, under what circumstances and how?' or more fully "what works, how, why, for whom, *to what extent* [emphasis added] and in what circumstances, in what respect and over what duration?" (Wong, Greenhalgh, Westhorp, Buckingham & Pawson, 2013). Realist evaluators are concerned with program theory and Context-Mechanism-Outcome configurations rather than entire programs. Realists investigate the contexts in which mechanisms – lying within people and society or introduced in interventions – fire to generate observable outcomes. To realists, programs work due to their effects on mechanisms or context. When considering the 'what', the realist position is that programs are not stable, single entities emitting some steady force for change (Pawson 2013, p. 48). "Mechanisms are the agents of change. They describe how the resources embedded in a programme influence the reasoning and ultimately behaviour of programme subjects" (Pawson, 2013, p. 115). Programs may also work because they address the context, or social structures that affect mechanisms (Astbury & Leeuw, 2010, p. 370).

This article aims to demonstrate how a policy maker, who might by sympathetic to a realist approach for understanding programs, but who will ordinarily look to an RCT for evidence of outcomes, can instead use an experimental design with a focus on program theory, to provide evidence about the value of an intervention, for whom, under what circumstances and to what extent.

While taking a realist approach, the article recognises the need for science to include the construction of falsifiable theory put to the test by experimentation (Popper, 2005) and by extension, for a realist scientific evaluation to test CMOs. This is not simply about sub-group analysis, identifying who seemed to benefit most and least in a pattern of results – either in an experimental data or a realist intra-program analysis. It is the view of this paper that a scientific, useful or portable realist CMO should be 'transfactual' that is, it should say something about the way the 'real' world operates outside a particular program or dataset.

This paper argues that experimental methods can be used to test theory, estimate the magnitude of an outcome in a hypothesised CMO configuration and assess the relative merit of an intervention for different target groups. The approach means shifting the

focus of experimental analysis away from the program or intervention towards program theory. Specifically it means making realist CMO configurations, rather than a program or intervention, the unit of analysis for experimentation.

While experimentation in science is much broader than the use of control groups, this is the most commonly used experimental method in the social sciences, and while it has limitations is the one used in this article. The particular method described used was a matched-pair, pre and post-test, control group design. This is by no means the only or best means of applying the general approach, but it was feasible in the evaluation where we sought to combine realist and experimental approaches to evaluation.

The key point of the article is that if a CMO is an important and useful description of the world, then – despite any shortcomings of experimental design using control groups – we should be able to test it and observe a regular outcome pattern in most instances when we observe a C (Context) and M (Mechanism) together – ideally with, but even if we don't have evidence of the M firing. If we do not observe a regular outcome pattern on a sufficient number of occasions, we may need to refine or abandon our hypothesised CMO.

The paper also addresses a key question often asked by those commissioning evaluations of public policy, one that we expect will still be asked even if realists are successful in shifting policy makers from focusing on 'what programs work?', to asking 'what works for whom, and in what circumstances?'. This is the question of 'how big was the effect?', or 'how big an effect can we expect if we leverage this CMO in the future?'

The realist case against excluding context from experimental evaluation of interventions into complex social systems

The question for experimental design as applied to the evaluation of public policy and programs is often of the nature 'what works and to what extent?' The unit of analysis is most often the program. Typical uses of experimental design for evaluation of interventions into social systems involve a treatment group, which receives an intervention, and a control group, which is supposed to be equal to the treatment group in all factors except exposure to the intervention. The means for ensuring the equivalence of treatment and control groups is either the random allocation of a sufficient sample of participants or the matching of participants in both treatment and control groups through a quasi-experimental method, such as propensity score matching. The outcome is then measured as the difference between treatment and control groups on some key variable of interest after the intervention. Since other factors have been 'controlled for', this outcome is attributed to the impact of the intervention.

Context is often treated as a confounding variable, and attempts are made to 'control for' the impact of context on outcomes through design or statistical analysis. More recently, proponents of experimental design have identified context as something to

consider when judging whether the results of a trial are applicable to populations other than those participating in the trial. However, the focus of analysis is still on the context-free impact of a mechanism or intervention. For example, in a collection of works for translating health research to public policy, reference is made to "understanding context-based factors that will have an impact on the success of interventions" (Wethington & Pillemer, 2012, p. 4) as an opportunity to contribute to translational research. The implication is clear; context is a factor that affects all aspects of an intervention rather being relevant for specific mechanisms. In the same volume, Evans conceives context as something to be addressed outside the experiment, rather than a critical part of what is being put to the test. He claims "social and behavioural science can be used to provide descriptive information on the community including family, social or political context in which interventions or policies are taking place, shedding light on the contexts in which desired changes are more likely to occur and on instances in which change is more difficult" (2012, p. 28).

Realist evaluators however, view mechanism and context as inexorably intertwined: controlling for the impact of context – as experimental designs often attempt to do – a is neither useful nor possible. The firing of a mechanism is completely dependent on context. To use a famous realist example, gunpowder does not fire when it is wet. Valid experiments are never easy to conduct, and experiments have received strong but sound criticism from realists as being ill-equipped to measure changes in complex social systems. Even if random allocation could achieve equivalent groups (at least on factors deemed important to achieving outcomes) prior to an intervention, the reality of ever-changing conditions, both within and between people, and the fact that context is part of what causes an outcome, mean an RCT that seeks to control context will often miss exactly what should be understood. The external validity of experiments will be limited as long as the unit of analysis remains the program and will be problematic whenever researchers attempt to control for the effects of context, rather than embrace context as determining whether mechanisms are activated and generate outcomes.

The common role for observation in realist and empirical social science

As with science generally, both experimental and realist approaches to evaluation rely on empirical observation. In the positivism influencing much experimental design in social science, knowledge is limited to observations of events. Realism posits that a deeper reality is knowable even if there is no such thing as final truth or knowledge.

In his 1975 landmark publication, *A Realist Theory of Science*, the realist philosopher Bhaskar (2008) argued against acceptance of this limited positivist conception of the world; "because it must be assumed, if experimental activity is to be rendered intelligible, that natural mechanisms endure and act outside the conditions that enable us to identify them" (p. 2). That is, we would not do experiments if we didn't think they told us something about the world outside the experiment.

Bhaskar argued the world is stratified into the domains of the real, the actual and the observable. The real is what exists, the structures and mechanisms that interact

regardless of whether they manifest into actual entities or events, and regardless of whether we observe these or not. For Bhaskar, (2008) “the real basis of causal laws are provided by the generative mechanisms of nature… [and these are] “nothing other than the ways of acting of things”… “tendances”… [or] “powers and liabilities of a thing which may be exercised without being manifest in any particular outcome” (p. 3). This means that in the complexity of the everyday world countless mechanisms are interacting in countless contexts, with the potential to lead to actual events that we sometimes observe. But real mechanisms exist even if they aren’t obvious, or act with consistent outcomes in actual (or factual) events that we observe – the real “exist independently of and are often out of phase with the actual pattern of events” (p. 2). In other words mechanisms are transfactual; they exist at a deeper level but give rise to everyday experience because “their activities are continuous and invariant, stemming from their relatively enduring properties and powers, despite their outcomes displaying variability in open systems” (Archer, 1998, p. 195). For example, positivists may seek to understand racism by measuring the regularity of ethnic minorities being passed over for jobs. For realists, the goal is to understand the mechanism of racism, which, although invisible, really exists even if it can only be observed during actual events.

As the domain of the real is not directly accessible to observation, social scientists are required to develop their understanding of reality though the observation of actual events, even though they are concerned with the underlying generative mechanisms of events, or abstractions such as the reasoning of program participants, that are not directly observable. Post-positivism may have somewhat bridged the gap between realism and positivism by accepting realist ontology, including Bhaskar’s argument about the intelligibility of experimentation, but methods of experimentation in program evaluation have not followed suit. Positivists tend to prefer to focus on the intervention as the unit of analysis and maintain fealty towards the ideal of invariances in experimental data – summed up in Hume’s famous phrase “the constant conjunction of events” (Bhaskar, 2008, p. 3). While there is acceptance of variation in data on outcomes in experimental evaluation due to the impossibility of controlled experiments and a reliance on randomised controlled trials, this is generally considered ‘noise’ that hides the true impact of an intervention rather than integral to the theory of how something works.

Making use of experimental designs in realist evaluation

Scientific enquiry often involves developing and testing theories using experiments. A true experimental design should test a hypothesis, not an intervention. While most program evaluations using experimental design seek to test whole programs or interventions, rather than program theory, this is a problem with the application of experimentation rather than experimentation per se. These types of evaluation are referred to disparagingly as ‘black box’ evaluation (Funnell & Rogers, 2011, p. 4). An experiment, used properly, provides a means of testing whether a theory can say something useful about the way an intervention works. Problems of external validity (i.e., how likely the result of an experiment is to apply in the real world) will occur whenever the program rather than theory is the unit of analysis.

Realist evaluators argue that it is impossible to isolate and measure the impact of an intervention, or of individual mechanisms not only because of the complexity of their interactions, but because it is the interaction of context and mechanism that generates outcomes. The unit of analysis is not the mechanism, or the context but the Context-Mechanism-Outcome configuration (Pawson & Tilley 1997 p. 217) [1]. In many cases, the complexity of social systems requires us to engage in developing middle-range theories as per Merton (1949):

> Middle-range theory is principally used in sociology to guide empirical inquiry. It is intermediate to general theories of social systems which are too remote from particular classes of social behaviour, organization, and change to account for what is observed and to those detailed orderly descriptions of particulars that are not generalized at all. Middle-range theory involves abstractions, of course, but they are close enough to observed data to be incorporated in propositions that permit empirical testing (p. 39).

Realists are concerned with middle-range theories and outcome regularities, demi-regularities or simply "demi-regs" (Pawson, 2010, p.185). But it is not enough for realists to simply hypothesise or develop more sophisticated theory "it must be possible for an empirical scientific system to be refuted by experience" (Popper, 2005, p. 18). Realists require a means by which Campbell's "mutually monitoring disputatious community of truth seekers" (Pawson, 2013, p. 192) can adjudicate disputes about the value of different interventions by testing hypothesised CMO configurations. This paper argues that realists of the Pawson and Tilley (1997) school, i.e., excluding critical realists, should aim to observe outcome patterns in contexts where mechanisms are hypothesised to fire by making predictions and testing them. If a theory cannot be tested in the observable world, it will struggle to be accepted as scientific. Interventions and their mechanisms may be difficult to define precisely; theories may be 'middle range'; and the observations may be patterns rather than constant conjunctions. However, conducting this work and gradually accumulating knowledge is all part of the slow, painstaking "informed guess work" of Popper's approach to science (Pawson, 2013, p.192).

Limits to intra-program comparison for testing theory

Testing hypotheses is not the same thing as looking at the pattern of results of an evaluation and making intra-program comparisons to develop theories about what works for whom, in what circumstances, and how. A dataset from a particular evaluation may be used to construct a theory of what 'worked' for whom under what circumstances, but not for testing a theory of what 'works' for whom in what circumstances – of transfactual CMO configurations. The danger of relying on

[1] This article like much of realist evaluation sidesteps the issue of the 'emergence' of outcomes from the interaction of psychological agency and/or sociological structure (Archer, 1998, p. 356) by seeking to engage our understanding at the point of interaction using CMO configurations as the unit of analysis.

data about what happened in a program is 'over-fitting' the data – "the most important scientific problem you've never heard of" (Silver, 2012 p 166). This happens when we seek to explain something by looking for patterns in a particular set of events that does not in fact explain anything about the underlying reality that caused them. Realist analysis that relies on fitting CMOs to data as well as experimental design that has insufficient attention to theory will fall short of a scientific means of measuring the impact of interventions into complex systems. As Nate Silver (2012), comments:

> What happens in systems with noisy data and underdeveloped theory – like earthquake prediction and parts of economics and political science – is a two-step process. First people start to mistake the noise for a signal. Second this noise pollutes journals, blogs and news accounts with false alarms, undermining good science and setting back our ability to understand how the system really works. (p. 162)

There is currently no perfect solution to the question of how to measure outcomes of interventions into complex social systems, but the imperative for realists to put their theories to the test is so strong that they should manage the deficits in experimental design instead of abandoning them. In any social experiment, there will be many things affecting outcomes in addition to a hypothesized CMO. However, if the CMO does explain something about the world, and it is of sufficient importance to be worthy of scientific study, then we should be able to observe patterns in outcome data as a result of an intervention which changes context, or introduces new reasoning or resources.

Realists could employ experimental designs as a means of testing theories and providing evidence of the demi-regular outcomes of a hypothesized CMO. Just as a large effect size can be identified in a small sample, a sufficiently useful and transfactual CMO will overcome the problems of dynamic systems and non-linear outcomes by being associated with demi-regular outcomes. If we do not observe a regular pattern of outcomes on a sufficient number of occasions, we may need to refine or abandon our hypothesised CMO as a useful concept for understanding the world and designing future interventions.

Experimental design for estimating effect sizes

Realist decision makers may be willing to accept the realist logic that interventions comprise mechanisms that work in different contexts. They may find a CMO built with intra-program comparisons compelling and useful for program design and targeting; and may be willing to accept that different interventions are required for different people. However, in order to make decisions about the relative merit and cost-effectiveness of the programs they administer, they will require an answer to the question: 'how substantial was an outcome when the mechanism fired?' An experimental design with a control group provides some ability to measure the independent impact of a mechanism firing in context apart from the impact of the countless other mechanisms affecting observed outcomes in open systems.

In the case discussed in this paper we attempted to control for the effect of mechanisms on outcomes by the use of a treatment and control group design. The difference from many forms of experimental design was avoiding randomisation to control for context. Randomisation does not allow the theoretically important contextual factors to be brought into the equation. Instead we used a matched-pair approach to bring context into the experiment. Both the target and the control groups were constructed to have similar initial conditions, to the extent possible, in terms of the contextual factors deemed important for firing the theorised mechanism. In this case, psychological wellbeing and resources were measured using psychometric scales. Crucially, we expected that change over the period of measurement (i.e., the school year) would be affected by many things outside the intervention, such as students' natural maturation and many other mechanisms at home and school. Our method allowed us to estimate how much of the change in wellbeing was due to the mentoring mechanism by focussing on the difference in the amount of change for treatment and control group students. Estimates of the size of an O in a CMO that do not seek to isolate the effects of different mechanisms on the O may mistake changes that result from many mechanisms as evidence about the particular mechanism within a CMO.

A realist quasi-experimental design for an evaluation of youth mentoring

This section demonstrates how we used a realist approach to experimental design for evaluating a youth mentoring program and estimating the size of an outcome that could be attributed to a mechanism within a CMO.

The mentoring program allocated community volunteers as mentors to students at risk of disengaging from primary or secondary school. The physical context in which the program was delivered was quite consistent: one hour once a week on school grounds. Generally speaking, the mentors did 'fun' things with their mentees – such as cooking and playing games – as a way of developing a trusting adult – student bond. The program aimed to increase students' psychological wellbeing and levels of important psychological resources, resilience, optimism, social skills, love of learning, as well as increase school attendance and reduce problematic behaviour.

The objectives of the evaluation were to measure the outcomes and identify how they could be maximised. One important question for policy makers was, 'when the type of mentoring we have on offer is provided to the type of students we think stand to benefit most, how much do they benefit?'

The evaluation used a realist synthesis of mentoring conducted by Ray Pawson (2004) and case studies to identify potential CMOs. These focused heavily on understanding what aspects (or mechanisms) of the mentoring program worked for which students and how. We only found one relevant CMO from this theory-building stage (see Figure 1), in part because the type of mentoring the program offered was limited to the mechanism of 'affective contacts' or developing emotional resources. Other forms of mentoring, such as advocacy, coaching or directing setting, were not included (Pawson, 2004).

Figure 1: The key Context Mechanism Outcome (CMO) configuration identified in the evaluation

Context: Students with low self-esteem/self-confidence or poor social skills or low resilience, often manifested as shyness or acting out in class i.e. low engagement in school.

Mechanism: Regular and consistent one-on-one time with a trusted and respected non-judgemental adult who engages in activities directed by the mentee. This activates empowerment and increased self-esteem, confidence and social skills.

Outcome: Increased feelings of self-worth, i.e., self-esteem (and sometimes increased resilience and optimism) leading to observations of greater self-confidence, (and sometimes improved peer relations) and greater engagement in the classroom.

We then used a quasi-experimental method of observing changes in pre and post-intervention measures of psychological wellbeing for students allocated to a mentor, compared to a control group of similar students without a mentor. To provide a control group for the evaluation, schools were asked to nominate about double the number of students who they thought could get a mentor. A mentor was allocated when one became available and was matched to a student in the nominated group with whom they seemed to be a good fit (not necessarily the student in greatest need). It was a relatively ad hoc process. In this kind of matched-pair design students who received a mentor were in the treatment group; those who did not, were in the control group.

Measurement of outcomes of youth mentoring

We sought to measure the outcomes of mechanisms in the contexts (i.e., students with a particular psychosocial profile) from which the theory-building stage of the project led us to expect to generate outcomes i.e., our CMO[2.] The approach was in essence very simple. We used statistically matched pairs to test whether those students with lower levels of self-esteem and poor social skills who were allocated a mentor achieved greater gains in psychological resources compared to the control group of students (with similar initial levels of these psychological resources) who did not get a mentor.

What we achieved was a measurement of the effect size when we expected a mechanism had fired. If we had looked at the effect size when a mechanism actually 'fired', it is likely the estimated effect would have been larger[3]. This approach is analogous to analysing outcome data by whether someone was allocated to get a treatment (effectiveness) or actually got the treatment (efficacy). Our approach was closer to this latter intention-to-treat analysis.

We used ANOVA and *t*-tests and Cohen's *d* (Cumming, 2012) to identify and measure the effect of the mentoring mechanism in context. The results were statistically significant and large. The effect size of mentoring for the students in the treatment group whom we expected to

[2] We were also required to measure the impact of the intervention as a whole. We found that while almost all mentors and students enjoyed participating and generally developed trusting and respectful relationships, and that all students improved on all measures used in the evaluation over the school year, as a cohort those who had a mentor did not improve more than control group students without a mentor.

[3] To increase the validity of the findings, we planned to observe cases where both mentor and mentee felt that mentoring had occurred as planned and actually helped the mentee (i.e., the mechanism fired), but the data was insufficient to allow us to match data collected from mentors with their mentee.

benefit relative to similar students in the control group was large. Cohen's *d* on the net differences between treatment and control students on psychological outcome measures ranged from 0.71 to 1.38 which was significant using a one tailed *t*-test[4]. The biggest limitation for the evaluation was the small sample size. Despite planning to have matched data from over 200 students, by the end we had 143 pre-treatment surveys, and were only able to match these with 93 student post-treatment surveys of which only 13 could be matched to the control group condition. Data from other sources[5] and the indicative statistical findings supported the hypothesized CMO, but a larger sample size would be required to provide sufficient occasions to observe demi-regularities[6].

We presented data using point estimates with probability estimates (i.e., *p*-values). In hindsight, it would have been better to provide point estimates with confidence intervals. Not only are confidence intervals recognised as the best approach to reporting statistics (Cumming, 2012); the approach fits with the realist task of identifying demi-regularities. Probabilities rather than precise measures, which are judged to be either significant or not significant, are suggestive of constant conjunctions and expectations for invariant outcomes.

The findings of the evaluation were that the mentoring intervention did not work for everyone, but it worked very well for a small subset of students, those with low self-esteem and poor social skills. This was because the form of mentoring available focused on providing students with emotional resources. So while mentoring was 'fun' for nearly all students (in part because mentoring involved activities directed by the student as an alternative to being in the classroom) the particular type of mentoring provided did not meet the needs of most students.

Limitations in the statistically matched-pairs method

This paper is about an approach rather than a specific method, yet there were two main limitations in our evaluation relative to the ideal of matched-pair treatment and control group design. First, the resilience outcomes we measured were not exactly the mechanisms that were hypothesised to be at work (self-esteem and self-efficacy), although they were closely related[7]. With a separate theory development and experimental design phase, we may have taken separate measures of the slightly different hypothesised psychological mechanisms and outcomes. However, the biggest challenge was doing both theory building and theory testing within a single project. We had some idea of the mechanisms of mentoring from the realist synthesis by Ray Pawson (2004). Interviews with frontline program stakeholders (mainly teachers and principals at 15 schools) addressed process issues and identified what aspects of mentoring they thought worked for which students and why. We used both sources to draft CMOs.

[4] Tests were significant with alpha set at 0.05, but the small sample size meant not all outcomes were statistically significant. However, it is uncertain whether the convention of setting alpha at 0.05 has as much relevance outside the laboratory for evaluations of complex social systems using mixed methods where only demi-regularities rather than constant conjunctions are expected to hold.

[5] Many teachers and school principals interviewed were 'lukewarm' in their support for the program and most had a number of examples where they believed mentoring did not lead to any change or benefit for a student but frequently reported that those that students with poor self-confidence or social skills benefited from the program.

[6] This paper is about an approach and a particular method we used to implement that approach – it does not make any claims about the effectiveness of mentoring as a result of the statistical data obtained.

[7] See the theory of core-self evaluations in Elliot, Kaliski, Burrus, & Roberts, (2012), p. 202.

Luckily, the time needed to measure outcomes (using a standard pre and post-intervention measurement for treatment and control group students) allowed us to continue to work on theory development. It was fortunate that some of the outcomes we sought to measure were also mechanisms. This meant data on these had already been collected in pre-tests. For example, self-esteem could be a mechanism (generating other wellbeing outcomes), a context (low or high self-esteem as a starting point), and an outcome (more or less).

Second, the experimental design would have been better had we – put the CMO to the test by only providing the intervention to those students whom the theory predicted stood to benefit, but we did not have a good theory about this until halfway through the project. It would also likely have been contested by teachers in the absence of 'evidence'. This limit to hypothesis testing raised a potential criticism that we were over-fitting the data, as might happen when developing CMOs using intra-program comparison data. What we achieved in this specific evaluation, as in many messy real world evaluations, fell short of the ideal. By using our statistically matched-pairs we made theory, rather than the program, the unit of analysis and we attempted to bring the context necessary for firing a mechanism into equation, while controlling for the effect of other mechanisms, to estimate the effect size of an outcome in a CMO configuration. The point of this paper is not that we achieved this flawlessly, but that we demonstrated how it may be achieved, in a way that is consistent both with realist theory and the principles of experimental design.

Implications for the mentoring program

The results of the evaluation suggested two main options for decision makers: develop the program so that different types of mentoring could be made available based on student needs, or target the program to students with low self-esteem and poor social skills. The first option would have been difficult as the program relied on local community volunteers for its supply of mentors, so the second was emphasised. In practice, this meant the program was only worth running in schools with a substantial number of students that fit the target group identified by the evaluation. Decentralisation of funding for student welfare programs meant school principals could decide whether or not to use their resources to fund the mentoring program or some other intervention that might better meet the needs of their students.

Potentially, with a measure of the effect size of an outcome from a CMO compared to the effect size of outcomes from other CMOs, realist cost benefit or cost effectiveness analysis may be possible. This might estimate which intervention is likely to generate the greatest benefits (or effect sizes) given the cost of the intervention and the context in which it is to be deployed.

Conclusion

Realist and experimental approaches to evaluation both involve theories about the world developed and tested through observation. When testing theories, empiricists often seek to use an experimental design that takes the impact of context out of the equation. Realists consider context as crucial to their theories and seek to observe contexts when mechanisms fire to generate outcomes. The problems of complexity and the threats to external validity of experiments in open systems are real. The practical difficulties of implementing a realist experimental design are significant. However, it is the position of this paper that if a realist CMO is worthy of scientific study, and is to inform decision making about the relative value of interventions, it should be tested experimentally and the magnitude or effect size of an outcome estimated.

This paper argues for and demonstrates a method of experimental design for testing CMO configurations. The approach makes the program theory rather than the program the unit of analysis. It brings the contextual factors hypothesised to be important for firing a mechanism into the effect size equation. A control group is used to isolate the effects of extraneous mechanisms and contexts on an outcome. It is argued that this approach is consistent with the underlying principles of realist and scientific evaluation and may facilitate more wide-spread recognition of the benefits to public policy of a realist approach to addressing questions about resource allocation.

References

Archer, M. (1998). Realism and Morphogenesis. In M. Archer, R. Bhaskar, A. Collier, T. Lawson & A. Norrie (1998). *Critical Realism Essential Readings.* London & New York: Routledge.

Astbury, B., & Leeuw, F. L. (2010). Unpacking Black Boxes: Mechanisms and Theory Building in Evaluation. *American Journal of Evaluation, 31(3),* 363-381.

Bhaskar, R. A. (2008). *A Realist Theory of Science.* London: Verso.

Cumming, G. (2012). *Understanding The New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*. New York: Routledge.

Elliott, D.C., Kaliski, P., Burrus, J., & Roberts, R. D. (2012). Adolescent Core Self-Evaluations. In S. Prince-Embury & D. H. Saklofske (Eds.), *Resilience in Children, Adolescents, and Adults: Translating Research into Practice.* New York: Springer Science & Business Media.

Evans, V.J. (2012). Translation in the social and Behavioural Sciences: Looking Back and Looking Forward. In Wethington E & Duniforn RE (Eds.) *Research for the Public Good: Applying the methods of translational research to improve human health and well-being.* Washington: American Psychological Association.

Funnel, S. & Rogers, P. (2011). *Purposeful Program Theory.* San Francisco: Jossey-Bass.

Haynes, L., Service, O., Goldacre B., & Torgerson, D. (2012). *Test, Learn, Adapt: Developing Public Policy with Randomised Controlled Trials.* Kew, London: UK Cabinet Office National Archives.

Merton, R,K. (1949). On Sociological Theories of the Middle Range, *Social Theory and Social Structure.* Simon& Schuster, NewYork: The FreePress.

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

Pawson, R. (2004). *Mentoring relationships: an explanatory review*. ESRC UK Centre for Evidence Based Policy and Practice: Working Paper 21.

Pawson, R. (2010). Middle Range Theory and Program Theory Evaluation: From Provenance to Practice. In J. Vaessen & F.L. Leeuw (Eds.), *Mind the Gap Perspectives on policy evaluation and the social sciences. Comparative Policy Evaluation,* Vol. 16, pp.171-202). New Brunswick & London: Transaction Publishers.

Pawson, R. (2013). *The Science of Evaluation: A Realist Manifesto*. London: Sage.

Popper, K. (2005). *The Logic of Scientific Discovery.* London and New York: Routledge.

Silver, N. (2012). *The Signal and the Noise: The Art and Science of Prediction.* London: Penguin.

Wethington, H. & Pillemer, K. (2012). Introduction: Translational research in the social and behavioural sciences. In Wethington E. & Duniforn RE (Eds.) *Research for the Public Good: Applying the methods of translational research to improve human health and well-being.* Washington: American Psychological Association.

Wong, G., Greenhalgh, T., Westhorp, G., Buckingham, J. & Pawson, R. (2013). RAMESES publication standards: realist syntheses. *BMC Med, 2013;* 11- 20 doi:10.1186/1741-7015-11-21.