A realist diagnostic workshop

Ray Pawson and Ana Manzano-Santaella

Evaluation 2012 18: 176
DOI: 10.1177/1356389012440912

Article

A realist diagnostic workshop

Evaluation
18(2) 176–191

DOI: 10.1177/1356389012440912
evi.sagepub.com
SAGE

Ray Pawson
University of Leeds, UK

Ana Manzano-Santaella
University of Leeds, UK

Abstract
The realist approach can now be said to be part of the repertoire of evaluation methods. There has been a corresponding shift in methodological focus. Polemical thrust and counter-thrust about the realist contribution as compared to that of other evaluative approaches such as randomized trials and meta-analysis has given way to closer examination of its practice 'on the ground'. This article seeks to make a contribution to the literature on how to conduct realist inquiry through a constructive critique of recently published 'realist evaluations'.

Keywords
context-mechanism-outcome configurations, explanation, multi-methods, realist evaluation

Introduction

This article was prompted by a remark made by a candidate for a post on a project requiring, as it was put in the person specification, 'experience in realist approaches to evaluation'. Sagely, he remarked that realist evaluation was in danger of becoming the 'new grounded theory'. He was referring, one supposes, to the overabundant and carefree use of a convenient methodological tag – one that bestows some contemporary vogue to the research, one that distances the writer from vile positivism and one, most significantly, that allows the researcher to pursue a blend of close empirical analysis combined with a modicum of careful theory development. The parallel is not exact of course. Grounded theory experienced a rather major hiccup when the two founding authors (Glaser and Strauss, 1967) fell out on finer points of method – thus leaving followers to choose between them (e.g. Walker and Myrick, 2006). We can report, by contrast, sweet harmony in the Pawson and Tilley camp.

The charge, however, remains. The years since the publication of *Realistic Evaluation* (Pawson and Tilley, 1997) and *Evidence-based Policy: A Realist Perspective* (Pawson, 2006) have seen the

Corresponding author:
Ray Pawson, School of Sociology and Social Policy, University of Leeds, Leeds LS2 9JT, UK.
Email: r.d.pawson@leeds.ac.uk

publication of more than 100 articles claiming to be realist evaluations and about a score of systematic reviews declaring themselves to be realist syntheses. A glaring question arises – are they really realist? It turns out that this is a tough question and arguably the wrong question to ask. It has always been stressed that realism is a general research strategy rather than a strict technical procedure (Pawson and Tilley, 2009: ch. 9). It has always been stressed that innovation in realist research design will be required to tackle a widening array of policies and programmes (Pawson, 2006: 93–9). It has always been stressed that this version of realism is Popperian and Campbellian in its philosophy of science and thus relishes the use of the brave conjecture and the application of judgement (Pawson et al., 2011). From these vantage points, it can be seen that realist inquiry is a broad and welcoming church.

The question, however, will still not go away. For a strategy to be regarded as paradigmatic, by definition, there has to be a family resemblance within the paradigm. While the core strategy is always capable of development under new challenges, realist inquiry has and must have unique and distinguishing family features. The article goes back to basics in documenting and explaining some of these meta-theoretical essentials. It does so, however, in a ruthlessly practical manner. Acknowledging that there is a core framework implies that out there in the field there are likely to be better and worse approximations. One of realist evaluation's tasks is to improve programmes by distinguishing between situations with effective and ineffective implementation. By the same logic, it follows that methodological refinement will be generated by examining applications of the method with exactly the same critical and differential focus.

This brings us more precisely to the task of the present article, which provides a close examination of some published examples of 'realist evaluation' in order to diagnose potential weaknesses and to consider how these interpretations of the approach might be strengthened. Readers will appreciate the present authors are treading on collegial eggshells here and so we stress that studies selected below are chosen with a diagnostic purpose – to improve the generic programme of realist research. What follows are *not* like the UK 'Office of Fair Trading' proclamations on counterfeit goods and the examples are *not* featured to name and shame particular authors. Rather, the purpose of the article is to clarify realist evaluation's core strategies and to elucidate its core terms by examining their commission and omission in a series of studies.

There is no such thing as the perfect inquiry, realist or otherwise. The case studies that follow are chosen to represent examples of common drawbacks, the difficulties that many authors have faced in trying to render realist principles into realist practice. Accordingly, the case studies are grouped to investigate three shortcomings: i) absence of an explanatory focus; ii) working in one data medium method rather than being multi-method; iii) failure to investigate contexts, mechanisms and outcomes *in configuration*. Note finally that we have investigated each case only against specific published outputs and our diagnosis, here apart, fails to mention one of the biggest drawback of all to realist ambitions, namely the journal requirements in many a field to publish in three or four thousand words. Little wonder that realist contributions fail to find room for all that occurs within the black box and in the contextual surrounds of an intervention.

Realism's explanatory focus

Realist evaluation has its own slogan – 'what works for whom in what circumstances'. Since it was coined in the first text, this phrase has gradually been embellished to capture the multiple, contingent outcomes of all interventions. A more meticulous, if less snappy, version thus goes: 'what is it about a programme that works for whom, in what circumstances, in what respects, over which

duration'. The most significant rendition, however, adds a vital realist signature – 'what works for whom in what circumstances . . . *and why'*. We know there will be a complex footprint of outcomes; the trick is to explain it. Why are the winners winners and why are the losers losers? Why does a programme work in Wigan on a wet Wednesday and why does it then fail in Truro on a thunderous Thursday?

Realist evaluation is avowedly theory-driven; it searches for and refines explanations of programme effectiveness. One can find, however, several self-professed realist studies that lack this essential process. Kazi et al. (2011) provide a typical example. The evaluation traces the outcome patterns associated with a care coordination programme called 'Wraparound'. The intervention is described as a 'strength-based, family-driven process that works to empower families and decrease or eliminate the need for service providers while increasing and maximizing families' connection and use of natural support' (2011: 59). The programme theory, as described, is about assisting families to identify their strengths as well as their needs and thus to engage a bespoke team from local agencies and their immediate community who will work collectively to address needs. The working hypothesis is a familiar one in the social care field, namely that the bridgehead to recovery should be assembled on the pre-existing stanchions of family potential. The actual implementation of Wraparound is not explained other than in passing descriptions noting that the requisite care develops over several months and in 'any place the family suggests' (2011: 59). In short, we sit squarely in realist territory here – long implementation chains, multiple and varied stakeholders, bottom-up ambitions, tailored and differential access to services.

Impact is assessed using the Child and Adolescent Functionality Assessment Scale (CAFAS). This is described by the authors in the following terms:

> used to assess a youth's functional impairment, rated as severe, moderate, mild or minimal/no impairment . . . a tool to determine day-to-day functioning that might be impacted by emotional, behavioral, psychological, psychiatric or substance use problems . . . a compilation of subscales: role performance (subdivided into school/work roles, home roles and community roles), behavior to others, moods/self harm (subdivided into moods/emotion and self-harmful behavior), substance use and thinking. (2011: 59)

Passing quickly over the problem of measuring baselines and outcomes at such high levels of aggregation (Pawson, 2002) we come to a key aspect of Kazi et al.'s article. As part of their work routine, Wraparound 'professionals' complete the CAFAS assessment for each participating youth on a quarterly basis (we also pass over the associated problem of 'experimenter effects' in measurement). These repeated measures provide the research team with a real time indicator of progress (or lack of it) through the period of the intervention. Rather than concentrating on the net performance of all participants, the research team apply realist logic, breaking down the analysis of outcomes using a variety of contextual and processual variables. And it is through this analysis that the realist motif, 'what is it about X that works for Y in what Z' comes into full force.

These contingencies are listed in great detail, of which four are reproduced to give a flavour of the main findings, the first in relation to the CAFAS school subscale, the second and third on the behaviour subscale, the fourth on the home subscale:

- those with greater impairment denoted by the total score on their baseline CAFAS measure improved at a greater rate than those with lower levels of impairment ($r = -.283$; $p < .05$; $n = 57$, power = .59);

- 85.7% ($n = 18$) of female youth improved in this outcome compared to 55.2% of males ($n = 16$) who improved ($r = .323$: $p < .05$, $n = 50$, power = .57);
- the mean age of those who did not improve was 15 (SD = 3.12, $n = 16$) compared to a mean of 16.93 for those who did improve (SD = 1.2234; $n = 35$);
- those who did not receive money for clothing and personal needs improved at greater rates (65.9%; $n = 29$) than those who received the service (25%; $n = 2$; $r = -.301$; $p < .05$, $n = 52$; power = .59).

A plethora of similar sub-group, subservice, subscale differences are noted on the basis of which, the authors claim: 'having real time access to this information is vital for the continuing evaluation of services, especially where immediate attention to significant patterns may lead to production of greater improvement in outcomes' (Kazi et al., 2011: 65). Once again a realist ambition is strongly echoed – the notion of strengthening implementation and improving the targeting of interventions on the basis of careful attention to outcome patterns. But is this particular analysis a sound enough basis on which to do so?

There are reasons to disbelieve. The first is that variations in sub-group success are potentially infinite. Clients may be subdivided on any of the familiar face-sheet variables but also by finer distinctions marking their family, neighbourhood, peer, criminal, cultural backgrounds and so on. Given that the scheme is adaptive, the service adaptations as delivered can also be measured in very many different ways – what particular ministrations are encountered, delivered in which way, by whom, over what duration, and so on? Finally, of course, there are outcomes and, as noted, although CAFAS is already a veritable jewel box of indicators, there are many other ways of testing impact through self-report, psychological tests, institutional records, performance indicators and so on.

Given that the potential permutations are inexhaustible, how can we know the research has latched onto the really significant outcome patterns? One way is by 'data dredging'. Kazi et al. (2011) employ this method using the (limited) data matrix manufactured by the project. The outcomes noted above are indeed the ones that pass sophisticated and time-honoured tests of statistical significance. Yet some standard drawbacks apply here (Lieberson, 1985). A significance test can only respond to the particular indicator chosen to measure a particular input or output. Choosing an alternative indicator may mean that the relationships falls in or out of statistical significance. In the present instance, before pronouncing on 'progress' or 'improvement' care must thus be taken to acknowledge that the particular measurement modality, 'professional opinion on Y', may not necessarily square with other potential measures of Y listed above. What is more, in all multivariate analysis the significance of any particular covariate depends on which other variables have been included and controlled for in the model. As we have seen above, the list of candidate variables is endless ($X_1 \ldots X_\infty$) and by including ($X_{17} \ldots X_{23}$) or excluding ($X_{31} \ldots X_{37}$) it is again possible that any particular covariate may fall in or out of significance (see Kazi et al., 2011: 61). In short, for the realist, variations in programme performance are crucial but outcome patterns considered *alone* are only surface 'markers' or 'traces' (Byrne, 2002: 32), the *potential* outward signals of inner workings of programme in a particular manifestation.

Their practical significance is only revealed by explanation building. The lack of the anchor role of explanation building can be seen immediately in examining the report's key findings noted above. We need to superimpose the why-question:

1. Why, for instance, do those with greater CAFAS-scored impairment at baseline improve their School Assessment more than those with lower recorded deficiencies? Could it be that

they receive more attention on the programme? Could it be that they are more appreciative of the scheme? Could it be because it is their very first time they have been so encouraged? Or, could it be a measurement artefact – regression toward the mean?

2. Why do females outstrip males in the practitioners' assessments of improvements in behaviour due to the programme? Might it be that girls find 'empowerment' less embarrassing than boys, are more practiced at 'talking things through', and are more willing to interact outside their immediate group of peers? And might programme practitioners respond more fulsomely to these points of potential. Or, might it be an artefact of the research act, girls being more guileful under observation and the scorers falling back on stereotypes of placid female demeanours?
3. Why do older participants outpace the young in their behavioural improvement? Might they have seen more of the consequences of continued delinquency? Might it be because of growing awareness that life's difficulties will multiply as they come to an end of their school career? Might it be that they are more mature in their responses to practitioners? Might they be better practiced in manipulating care workers and extracting services? Or, might there be a measurement artefact – with older subjects being more practiced at 'faking good'?
4. Why did the receipt of payments for clothing and personal needs lead to lack of progress in relationships in the home? Might it be that the incentives are often squandered – breaking the kernel of trust on which the programme is based? Might it be that the family is shamed by their failure to provide? Might the family be sufficiently dysfunctional that individual payments cause internal feuds? Or, might there be an artefact of programme implementation in that these payments came on tap at a place or a moment opposed by practitioners?

The point of these instant conjectures is to show that statistically significant relationships don't speak for themselves. They are capable of multiple explanations and sometimes contradictory explanations and sometimes perverse, artefactual explanations. Without knowing which explanation applies, it would be grossly premature to adapt or retarget the programme. Sadly, it is still necessary to point out that correlation does not equal causation and that this ancient maxim also applies to the 'forward conditional binary regression models' applied in this research. Correlations must not be mistaken for explanations because variables do not have causal powers. For instance, it is not the 'age' of these young peoples' bones, which acts to improve behaviour while on the programme. Age bestows the person with certain experiences and it is the store of preconceptions that they bring to a programme that lead them to interpret it and act on it in different ways. Similarly, it is not 'payments' that drive deterioration in behaviour. The payments are a 'resource' and such incentives can be invested or squandered according to the subjects' reasoning. In all cases, the outcome patterns come to be as they are because of the collective, constrained choices of all stakeholders. In all cases, investigation needs to understand these underlying mechanisms in order to capitalize on the gains accrued in charting the differential effectiveness of the intervention.

How might this be achieved? How can the study be nudged into being more firmly realist? While there is no exact protocol for doing so, this particular design might have been buttressed as follows:

1. Realist research is theory driven and in this study could have usefully begun in prior qualitative work eliciting, articulating and formalizing some hypotheses about why outcomes are so varied. It is the task of programme practitioners to guide all comers through every assumption and turn of an intervention. Typically, they have an abundance of expertise on

who prospers in relation to which programme feature and, crucially, *why this might be so*. This is intensely practical theorizing and it would, incidentally, knock the spots off the top-of-the-head conjectures listed above to account for the four outcomes (for an example of practitioner wisdom on such matters see Pawson and Tilley, 1997:108). This is not to say that these 'folk conjectures' are perfect but they can provide a legitimate focus for the investigation – an alternative to the happenstance of dredging through the infinite programme disparities.

2. A second strategy would be to provide theoretical focus and thus better focus to the outcome analysis. Rather than a statistical trawl to discover which sub-group or sub-service or sub-scale difference is significant, pursuing a theory can drive the analysis to look for further 'cross item' corroboration. If we begin, for instance, with a theory that a certain level of maturity is needed for a youth to benefit from a 'wraparound' of service provision, we might speculate that this might show up in a pattern where age relates to behavioural improvement (as in finding 3). But we might also expect that age (as a proxy for maturity) would also relate to the youth's ability to deal with more services. Is there further quantitative evidence to suggest that age correlates with increased access with the agency office, school, residential treatment facilities, family court, church, community groups? In short, realist analysis needs to account for networks of outcomes rather than provide a catalogue of discrete outcomes (see Trochim, 1985).
3. A third strategy would be to further deepen the working hypotheses by consulting those on the receiving end of programmes using respondent validation (Pope and Mays, 2006). The recommendation so far begins with practitioner folk theories and seeks to corroborate them in the rich evidence on outcome patterns. If the original hypothesis posits that sub-group X prospers (or fails) because of their preference set Y, then it is also useful to ask members of sub-group X whether they too concur with the conjecture – do they recognize the broad theory as a description of their motivations? This evidence may be gleaned in follow-up qualitative research with selected sub-groups where participants are able to reflect back on their experiences of the programme.
4. A fourth strategy for making sense of the outcome patterns is to compare notes with (or formally synthesize) existing research on the same family of programmes. Although Wraparound is unique in time and place, its ambitions and structures are very well known in community social work. While such programmes never reproduce exactly, we already know quite a lot about for whom and in what circumstances and why its ingredients (mentoring, out-of-home placements, care coordination, etc.) work (Philip and Spratt, 2007).

In short, Kazi et al. are quite correct to pinpoint the discernment of rich outcome patterns as a cornerstone of realist evaluation but have failed to see that it is explanation that builds and sustains the pattern – a point made many years ago by the distinguished philosopher Abraham Kaplan: 'The pattern can be indefinitely filled in and extended: as we obtain more and more knowledge it continues to fall into place in this pattern and the pattern itself has a place in a larger whole' (Kaplan, 1998 [1964]: 335).

Winners and … winners?

Long before realist evaluation entered the fray, there was always friction between exponents of outcome and process evaluation, which rested in turn on the old antagonism between quantitative and qualitative research. Those preferring quantitative evidence regarded qualitative data as dangerously

subjective. Those preferring qualitative evidence regarded quantitative data as providing crude oversimplifications of the human response to interventions. Readers will remember those days and perhaps regard them as a thing of the past given that the current orthodoxy, realist evaluation included, recommends a multi-method approach. As a first approximation one can say that mining mechanisms requires qualitative evidence, observing outcomes is quantitative, and that canvassing contexts requires comparative and sometimes historical data. The requisite balance, however, is precarious and there are 'realist' studies that attempt to cover all angles in an essentially descriptive and thus qualitative manner. Artificial results often follow.

The ensuing difficulty has long been dubbed as the tendency to produce 'good news' stories. In old parlance, the problem involved authors of rich, qualitative accounts of the participants' positive interpretations of a programme going on to proclaim that it 'works' (and should be extended, funded further, etc.) without the benefit of any quantitative data on whether behavioural outcomes had actually changed.

Under the new species of 'qualitative realism' the embellishment is more subtle – the careful elaboration of how a programme may work carrying over into assertions that it has worked. An example is a study by Priest (2006) of a community capacity building programme, *Motor Magic*, seeking to address motor and sensory impairments in preschool children. The programme 'uses a setting approach within a kindergarten environment, aiming to provide easy access to occupational therapy for children and to maximize opportunities to engage with, and build the capacities of parents and kindergarten staff to support those children' (2006: 221). Again we find ourselves with a multi-objective, multi-component, multi-stakeholder programme, which Priest approaches in the realist manner beginning by breaking *Motor Magic* down in its programme theories. A dozen strategies are discovered within the black box – embedding occupational therapy classes within the standard curriculum, parents observing and participating in specific activities, formal training in therapy for kindergarten staff, etc.

Next, as music to realist ears, each strategy is then unpacked in the form of CMO conjectures (Pawson and Tilley, 1997: ch. 4), of which we paraphrase a couple of specific examples:

- the strategy of grouping children together with similar needs, is hypothesized to work best for 'children with language as well as fine motor difficulties' (Context) by allowing children 'to watch and copy others to develop their own fine motor skills' (Mechanism) with the result that there is increased willingness to attempt new activities at kindergarten and home, increased participation in fine motor activities at home, etc. (Outcomes). Priest (2006: 226)
- the strategy of 'parental inclusion in activities using their non-dominant hand' is said to work best for 'parents with limited understanding of their child's particular developmental needs' (Context) by generating 'more positive attitudes and increased pleasure and appreciation of their child' (Mechanism), with the result that there is an improved relationship with the child (Outcome). Priest (2006: 229)

Now we come to the matter of the empirical corroboration of the many hypotheses. Two focus groups were conducted with kindergarten staff and parents, led by an external facilitator (not the evaluator) exploring the various programme theories. Interviews were audiotaped, transcribed and analysed using the CMO hypotheses as the thematic frames. The methodological problem, in short, is that this same body of qualitative evidence is made to speak to the Cs, the Ms and the Os.

In the case of mechanisms, unsurprisingly, some compelling data is unearthed. Mechanisms are embodied in the subjects' reasoning and they are best investigated therein. The classic analytic device chosen is to quote passage after passage in which parents and staff articulate how they have

interpreted and acted upon the resources provided by *Motor Magic*. To follow just one example, one of the parents recounts the following experience of participating in motor skills activities:

> We were made to cut with the opposite, our left hand or right hand regardless . . . and write our names . . .and that really was an eye-opener for me to show how difficult for me it was as an adult, but as a child having to do these things right from scratch and not knowing how. That was a real eye-opener. (Priest, 2006: 228)

Programme mechanisms change minds. They open eyes. And such close qualitative research is how to go about uncovering them.

In the case of charting outcomes and contexts, other forms of data have more authority and qualitative data is stretched to breaking point in being made to measure and compare. In respect of the outcomes, the study does great initial service in hypothesizing and charting the wide range of benefits that might flow (willingness to participate, improved relationships, readiness for school, behavioural change, etc.). But for each output or outcome, the evidence is compiled under exactly the same narrative strategy of reproducing statements from proud parent or positive practitioner: 'I think that his self-esteem, his confidence that all just grew and he was a completely different little boy' (Priest, 2006: 223). And no doubt *he* was. But setting aside all problems to do with selectivity, social desirability effects, chatty bias, researcher partisanship, and so on, the problem is that hand-picked, personalized description of outcomes cannot reveal collective outcomes patterns. Realist evaluation presupposes pattern. There will be winners . . . and losers.

Since the task of the present article is to refine the realist approach we move to remedial ideas, which in this case are straightforward. Realist research works by explaining outcome patterns and these cannot be determined though anecdotal remarks (on the part of subjects) or wishful thinking (on the part of evaluators). Outcomes should be carefully conceptualized and indicators thought through; baselines should be established; before-and-after measures should be plotted; complete cohorts of subjects should be followed. None of these requirements necessitate a fundamental reorientation of research strategy. Many of them are also the prerequisites of good administrative data, which can often be harnessed in the close confines of case study research. The key point is to address theory and if, to repeat the examples above, the theory says that there will be an 'increased participation in fine motor activities' or 'an improved relationship with the child' then such increases or improvements should be monitored *and* apportioned.

We must make it clear that Priest's study bears all the hallmarks of a preliminary inquiry and doubtless the author would regard it as a modestly funded pilot, designed to raise explanatory hypotheses. In its research strategy, however, it does typify one mistaken interpretation of realist evaluation – as just another form of qualitative inquiry, opposed to positivism and designed to penetrate the intervention black box. The core realist intention has always been to complement the measurement of change by understanding how it is generated (Archer, 1998). By throwing light into the black box we understand the transformation of input to output.

It is interesting to note that Priest's inquiry is almost the mirror image of our first case study (Kazi et al., 2011). There, outcome data proliferated but without theory or qualitative evidence. Both of the latter appear in study two – in the complete absence of outcome data. Put the two strategies together and realist evaluation begins to be realized.

Configurations not catalogues

The most unlovely term in realist terminology is the 'context, mechanism, outcome configuration' – the CMOc. It is an ugly circumlocution but it is there for a purpose. The phrase attempts to convey

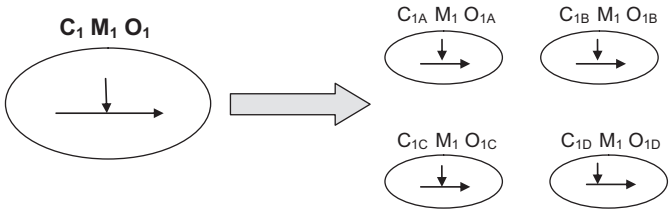

Figure 1. CMO before and after testing

the idea that evaluation tests programme theories and to do so the theory must be cast as an *if-then proposition*. The idea is to render the programme theory into its constituent and interconnected elements. In plainer, if more elongated prose, a CMOc is a hypothesis that the programme works (O) because of the action of some underlying mechanisms (M), which only comes into operation in particular contexts (C). *If* the right processes operate in the right conditions *then* the programme will prevail. To emphasize the causal and conditional nature of this conjecture the idea was presented formulaically as the 'equation': C + M = O (sometimes better rendered C + M → O). The action of a particular mechanism in a particular context will generate a particular outcome pattern.

As such, a CMO is a proposition and testable one to boot. There are different ways to construct the test, though in essence a realist investigation will hypothesize, monitor and seek to explain how the same programme resource is interpreted and acted upon in different ways by different participants in different positions. As a simplified example we depict this process of investigation in Figure 1. A programme based on, say, payments, loans, grants or giveaways may have started with the elemental proposition that the incentive (M_1) will encourage participants (C_1) to change their behaviour (O_1). Some initial conjecture would soon lead to a refinement, for instance that the same incentive (M_1) may be squandered (O_{1A}) by the rich (C_{1A}) and seized upon (O_{1B}) by the needy (C_{1B}). A more sophisticated theory and a tougher test would, of course, be provided by contemplating further sub-groups who might use the incentive in further ways, leading to the discovery of many other configurations as in Figure 1. In short, and in the familiar jargon, any starting CMO configuration is merely the beginning of the journey explaining what it is about the programme that works for whom and in what circumstances.

This propositional and proposition-building function of the CMO has not always been fully understood. The problem is that programmes never offer up a single theory. In realist terminology, there will always be multiple Ms – a proliferation of ideas within a programme, creating different resources that trigger different reactions among participants. There will always be multiple Cs – a huge range of individual, institutional and infrastructural features that condition the action of the assorted mechanisms. There will always be multiple Os – an extensive footprint of hits and misses, an uneven pattern of success and failure associated with the underlying causal dynamics.

In order to anticipate this inevitability, Pawson and Tilley (1997) devised 'the CMOc table' as a way of incorporating multiple hypotheses within any investigation. Table 1 illustrates the general format. If-then propositions are displayed *horizontally*. One of the ways in which the programme might work is described as M_1. It is supposed that this resource will only be acted upon by certain subjects in certain circumstances – described in shorthand in the table here as C_1 and O_1. It is supposed that this initial configuration will be put to test in the evaluation under the regime already described in Figure 1, identifying the many ensuing conditionalities, described above using the various subscripts C_{1X} M_1O_{1X}.

So much for the row one – the first configuration. The table then proceeds through a range of other CMO hypotheses, mindful that programme theorists always follow that chilling metaphor about skinning cats. The second row might consider quite a different aspect of intervention theory.

Table 1. Multiple CMOc propositions

Context	+	Mechanism	=	Outcomes
C_1	+	M_1	=	O_1
C_2	+	M_2	=	O_2
C_3	+	M_3	=	O_3
C_N	+	M_N	=	O_N

For illustration's sake let us imagine a youth job creation programme that combines incentive (M_1) payments with mentoring (M_2). Hypotheses are then built around this second configuration $C_2M_2O_2$, which assumes the ensuing discovery of a multiplicity of different responses to wise counselling C_{2X} M_2O_{2X}. The process continues, say, by considering the third, fourth and fifth means through which the programme may work. And so the hypotheses multiply until the researcher reaches exhaustion or the limits of funding. This basic table can be elaborated in many ways and at various levels of detail. It can also be used to describe CMO findings as well as CMO hypothesis. But in all cases they are CMO *configurations*.

Alas, researchers have found another way to interpret and produce these tables. Given the endless complexity of programmes and the situations in which they are embedded, it is a task in itself to contemplate the very many ways in which change might be engendered, the multiple constituencies of stakeholders and their myriad responses. Several researchers have thus taken the realist task to be the enumeration of the explanatory ingredients. One can propel investigation *down* the columns (especially if they are presented with gridlines). In so doing, the explanatory elements become atomized and disconnected. CMO configurations become unconfigured and transform into CMO catalogues. We examine two examples, each providing illustrations of the drawbacks and their consequences.

Our first sighting comes in a pioneering article by Ho (1999). It aims to provide a critique of the then limited and faltering approaches to evaluating urban regeneration programmes. Assuredly, the author dismantles the existing approach, which amounts to a 'stocktaking' of programme outputs (1999: 423). Community regeneration programmes are renowned for their complexity and output monitoring alone leaves us without a clue about which of the many community activities has affected change. The article thus goes on to propose a realist research agenda, which is constructed and labelled as a 'suggested framework for designing the context-mechanism-outcome configurations' (reproduced here as Table 2).

In short, this particular CMO table is intended for use in hypotheses-making. Our argument is that in this particular format it has insufficient purchase as an effective design device. The first problem is that the catalogues comprising each column, while already substantial, are in actuality unstoppable. For instance, in thinking through the contextual conditions that might shape the success of any intervention, Ho, wisely, goes beyond an understanding that context means the immediate locality. The fact that we are researching the Grimethorpe Estate in Sinkton does not tell us what it is about that place that might condition the effectiveness of the interventions. So indeed it might be Ho's 'local economic conditions' but such a list could easily extend beyond the items presented and include other factors such as transport access to places of work and skill profiles of the locals, and beyond that to other conditions such as designated regional support status and national economic well being (growth or recession). Still other social and cultural factors such as racial and ethnic balance might also propel or impede programme progress. The presence of other welfare interventions in the community (highly likely) and the previous regeneration programmes (quite likely) will also condition the success of any present effort.

Context		Programme		Outcome
Locality	SELECTION OF PROGRAMME	Economic	IDENTIFYING MECHANISMS	Employment
within conurbation		*enterprise support*		*new jobs*
near conurbation		*venture capital*		*protecting existing jobs*
deprivation in wider areas		*industrial grants*		*self-employment*
pocket of deprivation		*business loans*		who gets these jobs?
Local economy		Training/education		Securing jobs
declining industries		*pre-vocational training*		*full-time*
mine closures		*access course for HE*		*part-time*
single large industry		*customized training*		*relative security*
fair spread of industries		*job hunting skills*		*contract*
investment hot spot		*provision of childcare*		*trainee*
struggling for investment		*job link schemes*		who gets what?
Level of deprivation		Physical		Improved environment
infrastructural provision		*infrastructural improvements*		*boost image*
state of the environment		*environmental improvements*		*improved infrastructure*
housing condition		*land reclamation*		*improved facilities*
severity of unemployment		*provision of new premises*		*access to facilities*
inadequacy of facilities		*housing refurbishment*		who benefits?
level of crime		*provision of new housing*		Improved housing condition
		security improvements		who benefits?
		installation of CCTV		
		improvement of facilities		
		Social		Improved quality of life
		community safety measures		*safer communities*
		focus groups (safety; health)		*better health awareness*
		support for community groups		*community support*
		Planning for Real		*increased participation*
		residents forum		who benefits?

Table 2. The realist approach to evaluating regeneration programmes – suggested framework for designing the CMO configurations

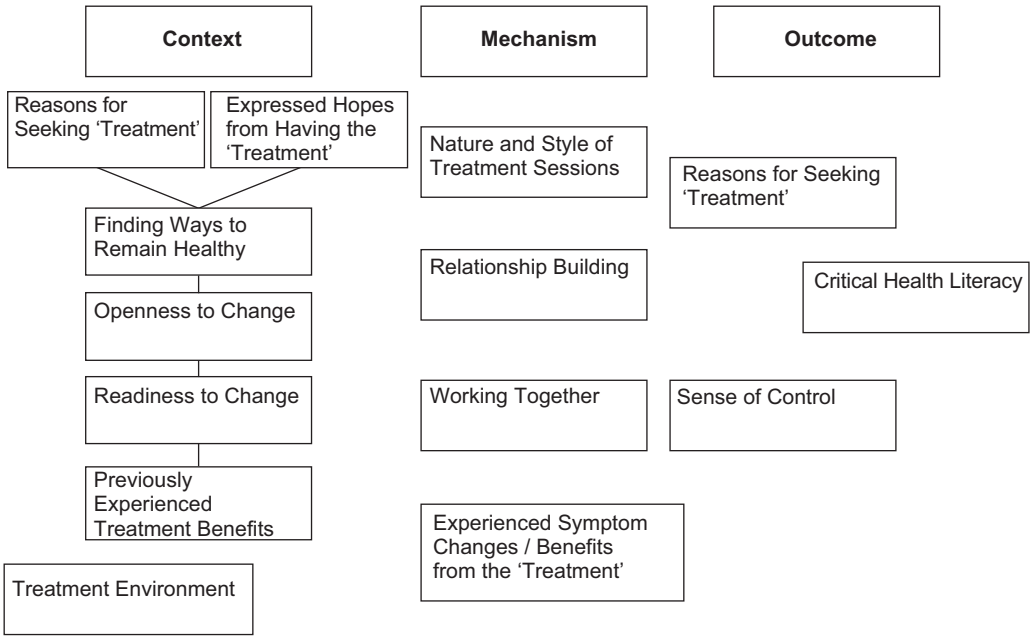

Figure 2. Unconfigured CMOs

The next problem with such 'ingredient listing' has become a classic realist headache – under which column does the hypothesized construct belong? In Ho's CMO table we note that the items 'infrastructural provision', 'infrastructural improvements' and 'improved infrastructure' run across the three columns. Presumably and fourthly, 'infrastructural provision' could also be considered an initial condition or baseline measure to the programme. Each designation is perfectly plausible: the infrastructural fabric of the estate could be so poor as to resist remedial action (C), that action could take the form of improvements to the housing stock and environmental facilities (M), one measure of the success of the intervention might be to compile a street-by-street, facility-by-facility inventory of changing conditions (O). The methodological remedy here, which we will explore more closely in the next example, is actually prefigured in the previous sentence – *which property falls under which category is determined by its explanatory role*.

A third and associated issue with the catalogue approach is the common confusion of programme measures with programme mechanisms. This may also have the ring of conceptual hair splitting but is in fact crucial to producing *propositional* CMOs. Ho lists some of the characteristic measures that are put in place in community regeneration – job creation schemes, community policing, business loans, CCTV installations, housing refurbishment, etc. Each of these may be considered an 'intervention' in itself, some ambiguity arising because these complex community initiatives, in common with much modern programming, assume that deep-seated problems only respond to broad-based remedies. Under realism, mechanisms are not regarded as the multiple components of combined interventions. Rather mechanisms penetrate to the layer beneath, attempting to explain how particular measures work. The nature of mechanisms is much discussed in the literature (Astbury and Leeuw, 2010) but for present purposes we retain Pawson and Tilley's (1997: 65) core notion that programme mechanisms capture the many different ways in which the resources on offer may impinge on the stakeholders' reasoning. To recall the original example, the

measure 'CCTV surveillance' may work through a range of different mechanisms – immediate arrest, improved detection, better deployment, increasing risk and deterrence, promoting natural surveillance, generating publicity and so on (1997: 78).

Ho, as a good realist, is aware of the distinction and inserts long curly brackets into Table 2, as well as some text explaining that a further framework is needed to accommodate stakeholder usage and choice. Accordingly, she descends once again the programme column hypothesizing, for instance, that training measures may work through increasing competitiveness in the labour market, physical improvements may increase, boost the image of the area for potential investors, and so on. While all of this reasoning is perfectly plausible, to our eyes it adds to the never-ending agglomeration of factors (problem one) that could potentially be addressed in an evaluation. Our notion of the function of CMO configurations as noted earlier is that they are rather narrow and limited hypotheses, which attempt to tease out specific causal pathways as pre-specified mechanisms acting in pre-specified contexts spill out into pre-specified and testable outcome patterns.

To make this significant point clearer, we move to a second example, Long's (2009) study of the potential of Shiatsu, a form of complementary and alternative medicine (CAM), in promoting health and well-being. This research also culminates in the production of a CMO table, which is reproduced here as Figure 2. In this case, the thinking behind the table emanates not from the arm-chair but rather from a particularly rich empirical study of key stakeholders. Long knows his mechanisms and contexts and gets right to the point in a longitudinal study uncovering what clients hope to get from Shiatsu and a survey of practitioners about what they considered essential in its delivery.

We provide no details of the fine textured map of Shiatsu practice that emerges from these inquiries (noting, however, that this is Long's main objective) and concentrate on the configuration as presented in Figure 2. Its role is to pull together a possible explanatory model for evaluating how CAM interventions might work. The layout is instructive. Both in the figure and the accompanying text, the narrative flows down the columns. Under 'context' the linked boxes identify particular client characteristics that mark them out as suitable cases for CAM treatment; they are seekers not sufferers. There is also a separate box for 'treatment environment', the characteristically interactive atmosphere that pervades CAM therapy. All of these features are well within the broad ambit of what realists refer to as context.

Next, we move down the mechanism column and things become murkier. The first box is labelled 'nature and style of treatment sessions'. There is no sign of client reasoning here, one of the defining features of a programme mechanism. Indeed, taken in isolation it reads very much like an aforementioned context, a pre-existing environment in which the intervention takes place. Then we have two boxes explaining how the treatment is taken on board – mechanisms to be sure. Then comes another perplexing box, 'benefits from the treatment', a phrase perhaps more redolent of outcomes.

Turning to outcomes proper, there are two boxes that interestingly point to a crucial idea that CAM clients pursue ends somewhat different from the norm. Confusingly, there is a further outcome box sharing exactly the same label as one of the initial contexts, 'reasons for seeking treatment'. Possibly the intention here is that initial expectations about the purpose of treatment are reinforced in the journey. It is impossible to say.

In short, in scanning the table as a whole one returns the predicament of unconfigured CMOs and the frequently asked question: 'I'm finding it hard to distinguish Cs from Ms from Os, what is the secret?' The answer is already prefigured. Long's CAM CMOs are untestable because they are composed of disconnected elements. While we lack the background to posit very specific

hypotheses in this field, it is possible to say how to render the constituent ideas into a 'model'. Basically the claim here is that clients with distinctive needs and characteristics (C) respond to the unique treatment relationships within CAM (M), so producing quite specific contributions to well-being (O). Tests imitate this propositional structure and can take several forms, the most basic of which: (i) follow clients who vary on the initial characteristics though a specific CAM regimes and trace differences in ensuing outcome patterns, (ii) follow the same cohort of CAM volunteers through the subtle differences in CAM relationship building and trace these through to reported outcomes. In all cases research interrogates the explanatory proposition.

Conclusions

Our summary remarks offer a brief recapitulation of the main methodological arguments before returning to comment on the very idea of a 'diagnostic review'. The first conclusion is that the phrase 'theory-driven' means what it says and that designs that attempt to utilize the realist explanatory apparatus without a prior grounding in programme theory will end with explanations that are ad hoc and piecemeal. A second tenet is discovered in realism's penchant for sitting in the middle, between positivism and constructivism, between process and outcome evaluation, between qualitative and quantitative research and so on. It should be stressed, however, that in the realm of research evidence, one obtains the best of both worlds by operating in both worlds. Process research cannot masquerade as outcome research and qualitative utterances cannot chart quantitative differences. The final conclusion addresses the tricky conflation of realist words and real things. Here we advise that programmes do not come in pre-ordained chucks called contexts, mechanisms and outcomes. Rather these terms take their meaning from their function in explanation and their role in testing those explanations. As wise mothers often insist – it is not the ingredients that make the dish but how they are brought together in the cooking process.

Some final caveats are now in order. The first is to reaffirm as strongly as possible the idea that the studies examined here are not completely fallacious. To repeat, there is no such thing as a perfect empirical study, realist or otherwise. A part of that pragmatic art is to apply compromises and cut the funding cloth as far as it will stretch. The immediate priorities of empirical research are to respond to the research brief, to deal with the given substantive issue, and to contribute to policy development – rather than to aim for methodological purity.

In the longer term, however, it is collective agreement on method that drives inquiry forward. That agreement is not set by dictate or decree. As Campbell famously argued: 'Somehow in the social system of science a systematic norm of distrust, combined with ambitiousness, leads people to monitor each other for improved validity. Organized distrust produces trustworthy reports' (Campbell and Russo, 1999: 143) In other words, methodological advance comes from a close scrutiny of each other's work and the ensuing debate.

It is with this in mind that we have named this piece a 'diagnostic workshop'. This is one of several instalments, in which a revolving list of realist concepts and realist evaluations and realist syntheses will come under scrutiny.[1] When it comes to methodology there is no such thing as the last word. All of the points made above are open to rebuttal and refinement. While it might be conceded that theory should be prioritized in realist research designs there are further issues about how programme theory should be sourced and selected (Weiss, 2000). While the battle cry that 'evidence should be a blend' might be conceded that, its exact balance still remains a battleground (Green et al., 1989). While it might be conceded that CMOs are configurational propositions, the matter of where to begin their construction is still open for discussion (e.g. are they really MCOs?).

The answers to these further questions are already 'out there' – it is simply a question of drawing them out in discussion and debate.

The basic ambition here is to prompt a rather unusual methodological phenomenon, which would imitate Campbell's idea of 'organised distrust' and would involve close monitoring of each other's work to improve its validity. Curiously, this rarely happens in the evaluation literature (see Gargani, 2010). There are, of course, very many examples of cut and thrust as paradigm wars make their periodic appearance in the journals. What we have in mind is something less vituperative – less about wrecking other methodologies and more about refining a preferred one. The idea is to take on studies in their own terms, over which task a commentator might spend several pages of close scrutiny, and in response to which the commented upon might yield here but defend there, and as a result of which the research community might find a better way forward.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Note

1. Another version of this article will appear in book-form (Pawson, 2013). It will also be published as a feature paper in the SAGE website http://www.methodspace.com. This offers a forum for reply, debate and discussion – an electronic version of 'organized distrust'. Methodological reflection on realist syntheses is also being taken care of by another collaborative project, which readers may join at http://www.jiscmail.ac.uk/RAMESES.

References

Archer M (1998) Realism and morphogenesis. In: Archer M, Bhaskar R, Collier A, Lawson T and Norrie A (eds) *Critical Realism: Essential Readings*. London: Routledge, 356–81.

Astbury B and Leeuw FL (2010) Unpacking black boxes: mechanisms and theory building in evaluation. *American Journal of Evaluation* 31(3): 363–81.

Byrne D (2002) *Interpreting Quantitative Data*. London: SAGE.

Campbell DT and Russo JM (eds) (1999) *Social Experimentation*. Thousand Oaks, CA: SAGE.

Gargani J (2010) A welcome change from debate to dialogue about causality. *American Journal of Evaluation* 31(1): 131–2.

Glaser BG and Strauss AL (1967) *The Discovery of Grounded Theory: Strategies for Qualitative Research.* Chicago, IL: Aldine.

Greene G, Caracelli V and Graham W (1989) Toward a conceptual framework for mixed-method evaluation. *Educational Evaluation and Policy Analysis* 11(3) 255–74.

Ho SY (1999) Evaluating urban regeneration programmes in Britain exploring the potential of the realist approach. *Evaluation* 5(4): 422–38.

Kaplan A (1998 [1964]) *The Conduct of Inquiry: Methodology for Behavioural Science*. New Brunswick, NJ: Transaction Publishers.

Kazi M, Pagkos B and Milch H (2011) Realist evaluation in wraparound: a new approach in social work evidence-based practice. *Research on Social Work Practice* 21(1): 57–64.

Lieberson S (1985) *Making It Count: The Improvement of Social Research and Theory*. Berkeley: University of California Press.

Long A (2009) The potential of complementary and alternative medicine in promoting well-being and critical health literacy: a prospective, observational study of shiatsu. *BMC Complementary and Alternative Medicine* **9**(19): 1–11.

Pawson R (2002) Evidence-based policy: in search of a method. *Evaluation* 8: 157–81.

Pawson R (2006) *Evidence-based Policy: A Realist Perspective.* London: SAGE.
Pawson R (2013 forthcoming) *The Science of Evaluation*. London: SAGE
Pawson R and Tilley N (1997) *Realistic Evaluation*. London: SAGE.
Pawson R and Tilley N (2009) Realist evaluation. In: Uwe H, Polutta A and Ziegler H (eds) *Evidence-based Practice: Modernising the Knowledge Base of Social Work?* Opladen and Farmington Hills, MI: Barbara Budrich, 151–80.
Pawson R, Wong G and Owen L (2011) Myths, facts and conditional truths: what is the evidence on the risks associated with smoking in cars carrying children? *Canadian Medical Association Journal* 182(8): 796–9.
Philip K and Spratt J (2007) A synthesis of published research on mentoring and befriending. URL: http://www.mandbf.org.uk/resources/research/
Pope C and Mays N (2006) *Qualitative Research in Healthcare*. Cambridge: Blackwell.
Priest N (2006) 'Motor Magic': evaluation of a community capacity-building approach to supporting the development of preschool children. *Australian Occupational Therapy Journal* 53(3): 220–32.
Trochim W (1985) Pattern matching, validity, and conceptualization in program evaluation. *Evaluation Review* 9(5): 575–604.
Walker D and Myrick F (2006) Grounded theory: an exploration of process and procedure. *Qualitative Health Research* 16(4): 547–59.
Weiss C (2000) Which links in which theories should we evaluate? In: Rogers P, Hacsi TA, Petrosino A and Huebner TA (eds) *Program Theory in Evaluation: Challenges and Opportunities* (New Directions for Evaluation no. 87) San Francisco, CA: Jossey Bass.

Ray Pawson is Professor of Social Research Methodology at the University of Leeds and author of *Realistic Evaluation* (with Nick Tilley), *Evidence-based Policy* and new work *The Science of Evaluation* to be published early next year. Please address correspondence to: School of Sociology and Social Policy, University of Leeds, Leeds LS2 9JT, UK. [email: r.d.pawson@leeds.ac.uk]

Ana Manzano-Santaella is a research fellow at the University of Leeds with expertise in health and social care programme evaluation and implementation. [email: A.Manzano-Santaella@leeds.ac.uk]