ONE

Precursors: Seven Pillars of Realist Wisdom

A couple of years ago Elliot Stern asked me to make a contribution to the journal *Evaluation* under the title 'From the library of Ray Pawson', with the aim of describing some of the intellectual precursors to my work. The intention was to provide a lighter-hearted coda to the other, more closely-textured papers evaluating water purification programmes in Swaziland, value-for-money local government schemes in Stoke-on-Trent, cheese production improvement in Switzerland, and so on. I took a little bit of persuading, fearing that it might read like a vanity publication, 'About Ray Pawson'. My feeling was that one's ideas should do the talking and that one's texts should speak for themselves. Better to watch Wayne Rooney play football than to have someone explain why he is good at it. Better to suffer prolonged, agonising torture than hear Wayne Rooney describe the influences on his game.

In the end humility faltered and the influences on *my* game were described (Pawson, 2011). Given that one of the aims of this book is to test the stamina of realist inquiry, it struck me that that little paper would make a fine starting point. I've adapted it by extending each pocket description as well as adding two further intellectual heroes. Rather than just describe their key ideas, I've also emphasised their value to evaluation research and evidence-based policy. With some partial exceptions, these are not the musings of evaluators. Most of these authors are writing to a philosophical remit and throughout the book I will call on the principles established in this chapter to provide the platform for a practical science of evaluation.

1. Bhaskar, R. (1978) *A Realist Theory of Science.* London: Verso

This, his first text, is on the philosophy of science and is unlike his remaining work which reaches into the social sciences and humanities. It gives pride of place in natural science explanation to the concept of the 'generative mechanism'. Physicists explain the relationships such as the gas laws through knowledge of the kinetics of molecular action. The attributes of compounds, such as gelignite's capacity to explode, is explained by their underlying chemical composition. Biologists explain evolutionary change through the mechanism of natural selection. In medicine, the 'mechanism of action' of a drug is what enables it to attack viruses, kill cancerous cells, heal bacterial infections, and so forth.

Bhaskar's moment of glory lies in showing that the laws of physics are not discovered through observational routines, nor through the mechanical application of measuring instruments. Rather, laboratory work attempts to reproduce a set of processes that we expect theoretically will give rise to an empirical regularity. Scientific experiments trigger a hypothesised mechanism in a known set of conditions in order to see if the expected uniformity comes to pass. Consider an early example, recounted by another realist (Harré, 1983) in a book on *The Great Scientific Experiments*. The experiment in question aimed to refine our understanding of the earth's geomagnetic field. It had been known (through observation) that magnetised compass needles always pointed in the same direction, originally considered to be towards 'Heaven' and then, more mundanely, to the 'North'. The prevailing theory explained this as a result of the north acting as a 'point attractive', which physically pulled magnetised objects in that direction. Norman's 1581 (Harré, 1983) experiment consisted of inserting a magnetised wire into a cork and suspending the floating compass in a beaker of water. As expected, the needle continued to point north. However, it also dipped in the direction of the North Pole but without descending to the bottom or moving to the north end of the beaker.

What this experiment does is to help confirm one theory at the expense of another. The conclusion drawn was that 'the direction is not produced by attraction but by a disposing and conversory power existing in the earth as a whole' (Harré, 1983: 53). The crucial experimental manipulation is the release of the compass needle from its fixed anchorage. Under this condition, we can then see clearly that it is not physically dragged off to the point attractant. Rather, its angle of rest means that the earth itself must have a magnetic field and that a free-floating compass simply aligns itself to the field. Our understanding of geomagnetism, it may be said, has advanced somewhat from the sixteenth century but this simple experiment has all the classic ingredients. Experiments are made by designing rather than observing a closed system, the design being informed by theory. The results of the manipulations are foreshadowed and interpreted by theories of underlying generative mechanisms which organise the observable properties.

There is a vast gulf between this account and the basic understanding of experimental manipulation at large in the world of randomised trials. Under the latter logic, laws are discovered through experimental manipulation and observation. The idea is to create (by randomisation or matching) two identical systems into one of which a new component is introduced. Observations are then made of outcome differences that occur between the experimental and control conditions and should a change occur, it is attributed to the one difference between them, namely the introduction of the experimental stimulus. The manipulation does not require any understanding of why the control situation behaves as it does and why the introduction of a new component might change it. The findings are expected to speak for themselves.

The message is clear. If evaluation is to follow the *Great Experiments* it would be wise for it to start with a theory of how the intervention affects the system into which it is introduced. As Bhaskar puts it, 'Theory without experiment is empty. Experiment without theory is blind' (1978, 191). From this pioneering study we glean a core concept, generative mechanism, and a pleasing motto to carry forward into the book. Note that following this groundbreaking work, Bhaskar's philosophical outpourings turned increasingly to the normative, the emancipatory, and, ironically, towards the point attractive heavens. Evaluation practitioners beware.

2. Archer, M. (1995) *Realist Social Theory.* Cambridge: Cambridge University Press

This is the first of several volumes whose task is to comprehend the nature of causality in the social world. Here we move from physical systems to social systems and from sub-atomic generative mechanisms to ... what? Archer's message is clear – social science should commence with an understanding of how people come to make choices, for their collective decision making constitutes the underlying mechanism that generates all social outcomes.

But there is a twist – society is made by, but never under the control of, human intentions. As she put it in a memorable phrase: 'Society is that which nobody wants, in the form they encounter, for it is an unintended consequence'. Our choices are set in a fascinating relay race. At any given time, choices are conditioned by pre-existing structures, institutions and opportunities. Those choices, once applied, then go on to remould a novel social structure, which in turn conditions a fresh round of choices for a slightly different cohort of choice makers. And so on. Society is thus in a state of permanent self-transformation (termed 'morphogenesis' by Archer). It is patterned and re-patterned by wilful action but without confirming anyone's wishes, even those of the most powerful.

This is the permanent state of society. This is how life goes on. There are major implications for evaluation research. The things we study – policies, programmes, interventions – are inserted into systems which are already fluid and changing. Interventions are often heralded as 'instruments for change' – but what they actually try to do, all they can do, is to change the course of change. This dynamic plays havoc with one of the time-honoured themes of evaluation research – counterfactual logic. Evaluation has traditionally been asked to pronounce on whether a programme makes a difference 'beyond that which would have happened anyway'. We always need to keep in mind that what would have happened anyway is change – unavoidable, unplanned, self-generated, morphogenetic change.

Evaluation will always struggle with the idea of an incessantly renewing world and it is worth setting down some of the initial implications. The first is to consider the nature of programmes. They are part of society and thus they too obey the iron law of self-transformation. Interventions will always mutate (thanks to sage practitioners, who always want to improve them) and can never be exactly reproduced (to the chagrin of dogmatic trialists who require them to be standardiseable). Then, there is the matter of what programmes do. Programmes seek to change the way that the subjects make choices. But, according to Archer's model, the collective choices of those so changed begin to make up a new social order. In short, programmes may well change the conditions that made them work in the first place and so can be expected to have a limited 'shelf life'.

Morphogenesis places a cap on the overall ambitions of evaluation. Programmes induce change but without conforming to the wishes of any particular stakeholder, even those of the most powerful. The changes generated can never be fully anticipated and are not entirely predictable. But this is no cause for hand-wringing; neither is social change haphazard and random. The social world stumbles between stability and change and it is the relatively enduring features that imprint a pattern. This world of 'demi-regularities' is the subject matter of realist evaluation and synthesis.

3. Elster, J. (2007) *Explaining Social Behaviour.* Cambridge: Cambridge University Press

Realist evaluation is a form of theory-driven evaluation. But its theories are not the highfalutin' theories of sociology, psychology and political science. Indeed, the term 'realistic' evaluation is sometimes substituted out of the desire to convey the idea that the fate of a programme lies in the everyday reasoning of its stakeholders. Good evaluations gain power for the simple reason that they capture the manner in which an awful lot of participants think. One might say that the basic currency is common-sense theory.

However, this should only be the starting point. The full explanatory sequence needs to be rooted in but not identical to everyday reasoning. In trying to describe the precise elbow room between social science and common sense one can do no better that to follow Elster's thinking. He has much else to say on the nuts and bolts of social explanation, but here we concentrate on that vital distinction, as mooted in the following:

> Much of science, including social science, tries to explain things we all know, but science can make a contribution by establishing that some of the things we all think we know simply are not so. In that case, social science may also explain why we think we know things that are not so, adding as it were a piece of knowledge to replace the one that has been taken away. (2007: 16)

Some of the things we all know, posits Elster, are encapsulated in the form of 'proverbial folk wisdom' and proverbs illustrate prettily how we must build on but also build beyond everyday understanding. So, if someone is stirred to observe that 'too many cooks spoil the broth' they are constructing a clever piece of everyday generative explanation about how over-staffed and chaotic work routines may lead people inadvertently to duplicate an action (the soup gets over-salted). Alas, posits Elster, proverbial reasoning has a tendency to mislead. The salty broth outcome is akin to but not consonant with another piece of metaphorical wisdom, namely that: ‘too many shepherds make a poor guard’. Here, over-staffing is such that people choose not to act under the assumption that someone else has already done the job. An even more contradictory proverb has it that ‘many hands make light work’. Forms of work organisation remain the explanatory mechanism. In this case, workers’ choices are preordained and fixed but their separate and individual functioning brings efficiency to a collective task. Finally, moving to tasks requiring uniform behaviour for all members of a group, we discover a further metaphorical twist to the proverbial repertoire on teamwork, namely the advice: ‘no member of a crew is praised for the individuality of his rowing’.

The point of comparing these everyday maxims is that they provide an evaluation challenge in miniature. They are all, so to speak, programme theories rooted in practitioner wisdom. They all point to outcomes, sometimes unintended, which are frequently discovered in collective work routines. They all feature mechanisms which tell us what it is about teamwork that generates a particular outcome. But no one proposition is universally correct; sometimes they are ‘so’ and sometimes they are ‘not so’. All depends on context. It is the realist evaluator’s task, and the added value of social science, to identify and explain the precise circumstances under which each theory holds.

4. Merton, R. (1967) *On Theoretical Sociology: Five essays old and new*. New York: Free Press

Evidence-based policy has become associated with systematic review methods for the soundest of reasons. Most programmes have a history and it makes sense to comb the historical records to see if we can discern reasons for success and failure. It always provides something of a shock in conducting such exercises to come upon the atomised and fragmented nature of programme-building. One digests report after report in which the same old programme theory is presented as an innovative intervention, with a shiny new acronym, aimed at a hitherto neglected social group, located in some previously overlooked corner of Never Never Land. Most evaluation research complies with this little conspiracy, tackling the one-off intervention using designs that start from scratch. One is led to wonder, under such a regime, whether lessons are learned, whether policy reflection has been deepened and whether programme implementation becomes more skilled.

But there is solace. There is a research domain, which is even longer in the tooth, and in which this problem is even more pressing and which foreshadows a solution. I refer to sociology, noting in passing that all of the social sciences trouble over whether they can be said to have progressed. Enter any sociology library and peer across the groaning shelves, enter cyberspace and download from all the countless sociology journals and similar questions are prompted. There are countless, separate inquiries but can they be said to cumulate? Has each author and each generation added to the wisdom of its forebears? Inquiry is older, but is it wiser? Answers to these questions seem to range from an unquestioning 'yes' (just look at all that work!) to a hostile 'no!' (why look for conflux when the job is social criticism?). To be sure, the issue of 'accumulation' raises a moot question but at least, for sociology, a solution has been mooted.

The blueprint for a progressive, accumulative social science has been long established. In one of his five essays, Merton puts forward the notion of middle-range theory, suggesting that we should produce explanations that: 'are sufficiently abstract to deal with different spheres of social behaviour and social structure, so that they transcend sheer description' (1967: 68). The key step comes with the ability to 'confederate' seemingly diverse empirical phenomena:

> An army private bucking for promotion may only in a narrow and superficial sense be regarded as engaging in behavior different from that of an immigrant assimilating the values of a native group, or of a lower-middle-class individual conforming to his conception of upper-middle-class patterns of behavior, or of a boy in a slum orienting himself to the values of a settlement house worker rather than the values of the street corner, or of a Bennington student abandoning the conservative beliefs of her parents to adopt the more liberal ideas of her college associates, or of a lower class Catholic departing from the pattern of his in-group by casting a Republican vote, or an eighteenth century French aristocrat aligning himself with a revolutionary group of the time. (1968: 332)

He suggests here that all of these seemingly diverse behaviours have a common thread. That dynamic is explained under an idea known as 'reference group theory'. This is based on the simple, abstract idea that people base their own actions on the standards of 'significant others'. In order to discern where an individual's life-chances lie one has a common investigative challenge – to figure out which is his/her relevant 'in-group' and 'out-group', how much she/he aspires to the in-group, and how high are the barriers forbidding in-group membership?

Evaluators would do well to seek confederation across their findings. The penny might then drop that their gleaming intervention is not new at all and will have been tried before – and that the place to start evaluation is with the well-travelled programme theory that underpins it. Available policy levers are not that numerous and so programme theories are repeated ad nauseam. The starting point is to consider much more tenaciously the similarities between seemingly diverse programmes – what do they hold in common?

5. Popper, K. (1992) *The Logic of Scientific Discovery.* London: Routledge

Having a background in social science methodology has made me very wary about strong claims for evidence. Social research is supremely difficult and prone to all kinds of error, mishap and bias. One consequence of this in the field of evaluation is the increasingly strident call for hierarchies of evidence, protocolised procedures, professional standards, quality appraisal systems and so forth. What this quest for technical purity forgets is that all scientific data is hedged with uncertainty, a point which is at the root of Popperian philosophy of science.

Popper preferred the term 'critical rationalism' to describe the considerable reach of his philosophical perspective. Here we pick up the 'post-empiricist' thread of the work. Like Bhaskar, he argued that scientific laws are not established in experiment and observation. For Popper such a viewpoint committed the error of induction, for no run of favourable data, however long and unbroken, is logically sufficient to establish the truth of an unrestricted generalisation. Black swans lurk in prey of the 'law' based on the million observations that swans are white.

For Popper, as with Bhaskar, it is our theories which make sense of observable regularities. But empirical evidence still plays a vital role in scientific research for it is capable of falsifying or limiting the scope of those theories. Accordingly, he moves away from the 'one hypothesis, one test at a time' view of scientific inquiry and regards it as a continuous or 'evolutionary' process. Scientists face a puzzling set of observational patterns; they apply their creative imagination by putting forward a bold set of conjectures to explain the apparent uniformities; they then test the theories in observation and measurement, the tests revealing more complex empirical work than first envisaged; some explanations are then preferred according to their ability to explain the patterns as well as the exceptions to the patterns; certain theories survive which are then put to further testing and development as new puzzling observations come to light. For Popper (1992: 94), as with Merton, science grows with the cumulation of explanation, rather than on the bedrock of observational facts:

> The empirical basis of objective science has thus nothing 'absolute' about it. Science does not rest upon rock-bottom. It is like a building erected on piles. The piles are driven down from above into the swamp, but not down to any natural or 'given' base; and when we cease our attempts to drive our piles into a deeper layer, it is not because we have reached firm ground. We simply stop when we are satisfied that they are firm enough to carry the structure, at least for the time being.

What is good enough for natural science is good enough for evidence-based policy, which comes with a frightening array of unanticipated swans – white, black and all shades of grey. Here too, 'evidence' does not come in finite chunks offering certainty and security to policy decisions. Programmes and interventions spring into life as ideas about how to change the world for the better. These ideas are complex and consist of whole chains of main and subsidiary propositions. The task of evaluation research is to articulate and refine those theories. The task of systematic review is to refine those refinements. But the process is continuous – for in a 'self-transforming' world there is always an emerging angle, a downturn in programme fortunes, a fresh policy challenge. Evidence-based policy will only mature when it is understood that it is a continuous, accumulative process in which the data pursues, but never quite draws level with, unfolding policy problems. Enlightened policies, like bridges over swampy waters, only hold 'for the time being'.

6. Campbell, D.T. (1988) *Methodology and Epistemology for Social Science: Collected Papers.* Chicago: University of Chicago Press (edited by S Overman)

Campbell is rightly venerated for his classic texts on quasi-experimentation, known fondly in the trade as the 'old testament' (Campbell and Stanley, 1966) and the 'new testament' (Cook and Campbell, 1979). These books devised research designs and statistical techniques to reduce threats to the validity of field experiments and they form the basis of all modern work in that domain. However, Campbell was also an eminent philosopher of science and laboured for over thirty years in developing an approach that he variously describes as 'evolutionary epistemology' and 'post-positivist, critical realism'. And it is this contribution that is represented here by a volume of his collected writings.

His name lives on in the evaluation community, being celebrated by a group of scholars attempting to organise systematic review methodology under the auspices of 'The Campbell Collaboration' (www.campbellcollaboration.org/). Somewhat mischievously, I want to suggest that Campbell would have had his doubts about membership. In particular, there are two Collaboration shibboleths that do not accord with the writings of Campbell the philosopher. The first is the insistence on 'procedural uniformity': the idea that in order to achieve objectivity and reproducibility reviews must be carried out in the same fashion to the same protocol. The second is the 'hierarchy of evidence', the concentration on evidence gleaned from Randomised Controlled Trials (RCTs), the low credit rating afforded to qualitative research, and the virtual detestation of local, tacit knowledge.

To advance my sceptical case I turn to Campbell's own words. Here is what he has to say on: i) objectivity and ii) qualitative method:

> The objectivity of physical science does not come from turning over the running of experiments to people who could not care less about the outcome, nor from having a separate staff to read the meters. It comes from a social process that can be called competitive cross-validation and from the fact that there are many independent decision makers capable of rerunning an experiment, at least in a theoretically essential form. The resulting dependability of reports ... comes from a social process rather than from the dependability of any single experimenter. 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. (1988: 302)

> Qualitative knowledge is absolutely essential as a prerequisite foundation for quantification in any science. Without competence at the qualitative level, one's computer printout is misleading or meaningless. We failed in our thinking about programme evaluation methods to emphasise the need for a qualitative

> context that could be depended upon. One example is the frequent separation of data collection, data analysis, and programme implementation that was once characteristic of Washington's funding of programs ... This easily lead to a gullible credulity about the numbers on the computer tape, with the analyst in total innocence about what was going on in the program implementation ... To rule out plausible hypotheses we need situation specific wisdom. The lack of this knowledge (whether it be called ethnography or program history or gossip) makes us incompetent estimators of programme impacts, turning out conclusions that are not only wrong, but often wrong in socially destructive ways. (1988: 366)

The implication for evaluation and systematic review could not be clearer. Here is a clarion call to scavenge for evidence of all forms, quantitative and qualitative, outcome and process, measurement and gossip! But then there is the glorious twist represented by the first quotation. However high this evidence is piled, it will not lead to objectivity. What counts are the hypotheses that drive us to the data and the inferences that are drawn from the data. In order to harden such inferences, Campbell argues that theories must be tested and tested again, sometimes to destruction and sometimes to live another day. Above all, we need to attend much more closely and collectively to the quality of the reasoning in research reports rather than look only to the quality of the data.

7. Rossi, P. (1987) 'The Iron Law of Evaluation and Other Metallic Rules', *Research in Social Problems and Public Policy*, 4(1): 3–30

Rossi has made many fine contributions that would grace any evaluation library. More obvious candidates for an accolade might be the two pioneering papers with Chen, which make the earliest claims for the utility of a theory-driven approach (Chen and Rossi, 1980; 1983). Another classic is the punctilious*, Money, Work and Crime*, which eats up 348 pages in evaluating a single programme (Rossi et al., 1980). This intervention, the 'transitional aid research project' (TARP), was based on the idea of providing released prisoners with small, limited term financial incentives to facilitate their adjustment to life beyond the prison wires. Early trials of the programme were highly promising, the revolving door of reincarceration turning significantly more slowly for the intervention recipients than for the unsupported control groups. But, as with many demonstration projects, disappointment followed – with a later, larger trial based in different penitentiaries failing to show any net impact.

Rossi's team had collected sufficient data to peer, nay pour, into the black box of this programme. The intervention is a simple incentive, the 'money' of Rossi's title. Now, as with all programme theories, incentives work through their perceived utility to the subject. In the case of TARP, aid could be used to support job search, namely 'work', or conversely, it could negate the immediate need to find paid employment and so initiate a return to old habits, namely 'crime'. Indeed, within limits, the payments could be used to support any chosen vice or virtue. In realist terms, the intervention triggers opposing mechanisms and it is the balance of choices in the population under study that determines the net outcome of the programme. Such balances can be expected to differ from instance to instance, trail to trial. Different mechanisms may, as here, cancel each other out.

Rossi came upon *countervailing mechanisms* in much of his evaluation research career and was thus inspired to compose the 'Metallic Laws of Evaluation'. The most tyrannical insists on the following:

> The Iron Law of Evaluation: The expected value of any net impact assessment of any large scale social program is zero.

Rossi's tongue was firmly in cheek in the naming of his laws. His brain was firmly engaged, however, for he insists, with the iron law, that programmes work only when implemented in a particular way and only when targeted at well-defined outcomes, for the right subjects, in appropriate circumstances. Why Rossi thinks that this formula impels us towards zero aggregate impact is that 'large scale programs' generally overreach themselves. In other words, a programme theory finds favour, sometimes on the back of good news from a pilot investigation, and a huff and puff of activity breaks out in its wake, encouraging it into the hands of inexperienced practitioners, and expanding its market to ill-defined outcomes, tougher subjects and inauspicious contexts. Few interventions can survive that journey.

Emboldened by this maverick paper, I conclude with a further decree:

> *A Golden Rule for Evaluators and Policymakers*: Instead of imagining your job is to choose the most effective interventions, better to follow the iron law and to treat a chosen programme as a blank canvas in which your task is to choose the best means for its targeting and implementation.

All good advice manuals should end at the magic number seven and so I cut short my tour of the library at this point. It goes without saying that many other volumes and many other authors could equally have taken pride of place. I am thinking, for instance, of the master of generative explanation, Raymond Boudon, whose sociological work is a model for evaluative inquiry, being a perfect amalgam of principle and practice. Fortunately, I have had the opportunity to say this elsewhere (Pawson, 2009a). Another unforgivable omission is Carol Weiss, also a founder of the theory-driven approach in evaluation and the scholar who has best explained its utilisation potential – the 'enlightenment approach'. No excuses here – other than that this entire book may be said to be given over to her question – 'which links in which theories should we evaluate?' (Weiss, 2000).

As explained, the real purpose of this chapter is to examine the infrastructure of a methodology and thus the real motive for the above selection is to demonstrate that realist evaluation and realist synthesis stand on the shoulders of giants.