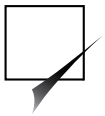

Article

Evaluation
2016, Vol. 22(3) 304–322

DOI: 10.1177/1356389016656518
evi.sagepub.com

EMMIE and engineering: What works as evidence to improve decisions?

Nick Tilley
Jill Dando Institute of Crime Science, University College London, UK

Abstract
While written by a proponent of realism, this article argues in favour of a pragmatic approach to evaluation. It argues that multiple sources of evidence collected using diverse research methods can be useful in conducting informative evaluations of programmes, practices and policies. It argues in particular that methods, even if their assumptions appear incommensurable with one another, should be chosen to meet the evidence needs of decision-makers. These evidence needs are captured in the acronym, EMMIE, which refers to Effect size, Mechanism, Moderator (or context), Implementation and Economic impact. Finally the article questions evidence hierarchies that are inspired by clinical trials, and suggests instead that, notwithstanding the clear differences in the physical and social worlds, engineering may provide a superior model for evaluators to try to emulate. And engineering is, above all, a pragmatic field.

Keywords
EMMIE, engineering, pragmatism, realism, realist

Introduction

This article embraces a realist orientation in evaluation, but its concerns are with the practical ways in which evaluation can best inform policy, practice and programme decision-making, even if this may sometimes involve departing from the realist ideal.

All programmes, policies and practices (PPPs) are concerned with improving existing states of affairs or patterns of behaviour or with averting unwanted future states of affairs or behaviours. Examples of states of affairs include morbidity and mortality rates, levels of unemployment, levels of literacy and numeracy, pollution levels, and the level and distribution of wealth.

Corresponding author:
Nick Tilley, JDI, Department of Security and Crime Science, University College London, 35 Tavistock Square, London WC1H 9EZ, UK.
Email: n.tilley@ucl.ac.uk

Examples of behaviours include crimes committed, taxes paid, discrimination exercised on the basis of race, gender or sexual orientation, types of food eaten, and school attendance.

Unfortunately PPPs often fail. King and Crewe (2013) have recently produced an entertaining if alarming catalogue of costly, major national policy flops in the UK during the quarter century leading up to 2010. These included, for example, the introduction of the poll tax, the formation of the Child Support Agency, the entry into and exit from the Exchange Rate Mechanism, the creation of the Millennium Dome, the operation of the Assets Recovery Agency, the single payments scheme for farmers, plans for National Identity Cards and sundry large-scale IT disasters. There is also no reason to believe that the UK is uniquely placed to produce policy and programme blunders.

Rational PPP decision-makers have an interest in reducing the risks of such failures and increasing the chances of success. Their wants lie in knowing what will produce which benefits at what costs. This knowledge should help them achieve better outcomes by informing the allocation of scarce resources in ways that will maximize utility. This explains the concern to create, assemble and use what works evidence.

With a view to improving the evidence base for decision-makers, in March 2013 Britain's Cabinet Office kick-started a series of 'What Works Centres' to draw together what works findings to 'help policymakers make informed judgments on investments in service that lead to impact and value for money for citizens' (https://www.gov.uk/guidance/what-works-network, accessed 28 May 2015; for the background, see also Her Majesty's Government 2012, 2013).

These What Works Centres covered, for example, health and social care, educational achievement, and crime reduction. Collectively the centres were intended to inform decision-making over £200 billion of public expenditure.

Each centre was charged with collating existing evidence, producing high-quality syntheses of evidence, assessing the effectiveness of policies and practices against an agreed set of outcomes, sharing findings in an accessible way and encouraging practitioners, commissioners and policymakers to use findings to inform their decisions.

One of the tasks of the What Works Centres has been to rate and rank available evidence for use by PPP decision-makers. This rating and ranking has become widespread and there is an appetite for criteria against which to score both primary studies and reviews of evaluation study findings. Those conventionally put at the top of the hierarchy are randomized controlled trials (RCTs), where threats to the internal validity of findings are allegedly eliminated by creating equivalent experimental and control groups by the blinded random allocation of subjects to one or the other, blinded allocation of 'real' treatment to the experimental but not control group, and blinded comparison of changes in the two groups to ascertain the effect size (if any) of the intervention provided in the experimental group.[1] This is deemed the 'gold standard'. Other studies are then ranked according to how well they approximate the RCT ideal. Best of all are repeated RCTs of the same treatment.[2] The results of suites of RCTs can be combined and the PPP presented with overall, highest and lowest effect sizes.[3] If the intervention is properly costed and the net benefits monetized then PPP decision-makers are provided with useful evidence on which to base their judgments. In this sense the evidence hierarchy seems to make good sense. PPP decision-makers want to know what to expect by way of impact, they want to know what it will cost and they want to know what the returns will be from different options.

The National Advisor for the network of What Works Centres, David Halpern, has stressed the virtues of the RCT as a vehicle for achieving improvement, suggesting that the successes

of the British cycling team provided a striking example of what could be achieved by a series of rigorous experiments testing the benefits from small changes (see Halpern, n.d.). Halpern also compared the rapid rate of expansion of health-focused RCTs with the relatively slow growth in those focused on social welfare, education, and crime and justice since 1900 (see also Syed, 2015). The What Works Centres have been focused on remedying the shortfall by assembling what there is already and helping to plug remaining gaps.

The next part of the article discusses the evidence needs of PPP decision-makers, which are not identical to their wants, and describes a method of rating that evidence: the 'EMMIE' scale. EMMIE was developed in the first instance specifically to apply to systematic reviews of evidence, hence to those forms of evidence that purport to be at the top of the evidence hierarchy, albeit that it is equally applicable to the assessment of individual evaluations aiming to inform PPP decision-makers. It will also be argued that it is relevant across the policy and practice waterfront, although it was initially devised specifically to relate to crime reduction.

For the purpose of this special issue of *Evaluation*, what is important is that EMMIE embraces a realist perspective with its stress on the formulation and testing of Context-Mechanism-Outcome Pattern Configurations (CMOCs), but also acknowledges the practical importance of 'black box'[4] RCTs and quasi-experiments that are rooted in traditional Humean conceptions of causality, which refer to constant conjunction.

The latter part of this article suggests that engineering may provide a better framework for doing and delivering EMMIE-informed evaluations for social programmes than clinical trials, which are often used as the inspiration for the RCT 'gold standard'.

What is common to the discussions both of EMMIE and engineering is the priority attached to trying to meet the practical needs of PPP decision-makers, even when this means incorporating research whose methodological assumptions appear to be incommensurable. In this sense the article aligns with Feyerabend's slogan, 'anything goes', provided it contributes to progress (Feyerabend, 1975).

EMMIE and meeting PPP decision-maker evidence needs

We turn now to PPP decision-maker evidence *needs* rather than *wants*, which have been referred to so far. Needs here refer to the kinds of consideration that PPP decision-makers must take into account in determining what they will and will not do.

It is difficult to see how evaluation evidence of any sort could speak to some matters decision-makers often need to consider. These include, for instance, values, ideology and public opinion. Decisions-makers have to take account of public reactions to what might be done, to fundamental values such as those to do with human rights or equity, to legal constraints on what must, can and cannot be done and to sundry political imperatives that face them. Choices are always made from options that are in practice available, albeit that the range of these may be contested. Several of the policy blunders referred to earlier, as identified by King and Crewe, may be attributable to ideology and political expediency, which would trump evidence-relevant considerations. This section, however, is concerned with PPP decision-maker needs that may be met by evaluation evidence.

The following specification of PPP evidence needs was developed as part of the Commissioned Partnership Programme in Support of the What Works Centre for Crime Reduction. This support programme was jointly funded by the ESRC and the College of Policing, which hosts the What Works Centre for Crime Reduction. One of the nine 'work

packages' required that, 'a research design … be proposed to develop a comparative labeling scheme, using a consistent evaluation standard to rate and rank the effectiveness of interventions and the overall cost-saving' (College of Policing and ESRC, 2013: 11). EMMIE comprises the labeling scheme that was created (Johnson et al, 2015).

'EMMIE' was devised to try to capture PPP decision-maker evidence needs. It is an acronym, where the two Es at either end refer to 'Effect Size' and 'Economic consequences'. These are clearly interlinked in the sense that the effect size provides one of the crucial elements for estimating the benefits achieved or expected (although estimating costs is tricky and monetizing benefits even more so, see Manning et al, 2016). The two Es refer to the evidence wants as well as needs of PPP decision-makers. There are good reasons to pay attention to them. Resources are always limited and decisions have to be made about how to allocate them. Rational PPP decision-makers need to take account of what effects can be expected and what the costs and benefits of those effects will be, preferably in monetized terms that permit comparisons across policy and practice domains. Moreover as PPPs are expanded or shrunk those trying to make decisions about them need to know what the marginal effects will be in terms of outcomes, costs and benefits. While evaluators, notably those well-versed in realism, may be correct in emphasizing inevitable uncertainties that relate to human intentionality, the openness of systems, and the diversity of changing contexts, it remains the case that for the rational PPP decision-maker best estimates of expected bottom-line net outcomes are important. This explains one of the attractions of the RCT and its closest practical counterparts: they offer an apparently hard-headed method for gauging outcomes.

RCTs and their quasi-experimental counterparts have therefore been commissioned and conducted to inform the evidence needs for the two Es. These methods of evaluation have, however, been little concerned with how interventions produce their effects. The intervention is thereby treated as a 'black box', where interest lies in what results from an intervention, rather than how its results are produced.

Coming to an informed view about expected effects and effect sizes and whether they warrant the expected costs of any PPP is, as indicated, sensible. Moreover, even if dependency on past performance involves a logical leap of faith, some inkling of what might follow on the basis of what has gone before is preferable to a guess or mere wishful thinking. Most previous ranking systems have been based on the internal validity of individual studies and reviews in their estimation of effect sizes and cost savings. The link between those and RCTs is obvious.

The practical benefits of RCTs are most obvious in clinical trials. Using four examples relating to the treatment of diseases of blood vessels and breast cancer, Peto et al. (1995) show that large-scale RCTs ('mega-trials') with heterogeneous samples that are allocated to different standard treatments (and/or no treatment) have been able reliably to identify small effects that hold the promise of informing treatment decisions that could prevent many deaths. Semmelweiss' trials on handwashing saved the lives of many women and their children (Semmelweiss, 1983 [1860]); trials for thalidomide might have averted the suffering of the offspring of women who took it at the crucial period of their pregnancy (Brynner and Stephens, 2001); and many women would have been spared the suffering from increasingly radical surgery performed on those with breast cancer in the often fruitless efforts to cut out enough of their bodies to prevent it from spreading (Mukherjee, 2011).

One problem with such trials is that they always relate to the specific populations from which allocation to treatment/non-treatment or treatment variation is made. Strictly they

cannot be assumed to go for other populations in other places or at other times (see Cartwright, 2007; Cartwright and Hardie, 2012). However, from time, space and population specific studies, '*can work*' and '*can cause harms'* inferences may sometimes validly be made (e.g. Semmelweis and handwashing, and Mukerjee and cancer), as can '*may not always work*' inferences (e.g. Brynner and Stephens for thalidomide as a treatment and Mukherjee for treatments of cancer). What cannot be concluded from a trial, however, is either that an intervention 'can never work' or that it 'will invariably work', even though answers to such questions may be highly desirable for PPP decision-makers. As for evidence counting against a treatment, it should be remembered that an RCT has not always been needed. It is sobering to remember that what did it for thalidomide was not a RCT but a clinician's observations of aberrant rates of severe abnormality among babies of mothers in his care and publication of a fifteen-line letter on this in *The Lancet* (McBride, 1961).

PPP decision-makers may want simple effect size measurements and associated cost–benefit estimations. Black box RCTs may seem to provide the answers. Unfortunately, however, although such RCTs are generally better than nothing in the evidence they furnish and can play an important part in improving some decisions about interventions, they are flawed as a sufficient basis for guidance.

Target heterogeneity means that the same intervention produces diverse and sometimes contradictory outcomes (see, for example, Davidoff, 2009; Kent et al, 2010; Peto et al, 1995; Rothwell, 2005a, 2005b[5]).[6] Indeed, physicians advocating evidence-based medicine have recently become concerned that the overall effects reported in trials and meta-analyses of findings may have led clinicians to disregard particular attributes of patients that indicate that the treatment that has been found to have an overall net benefit may not be appropriate for them and that informed clinical judgment is still needed (Greenhalgh et al, 2014). With an aging population where co-morbidity and multiple, possibly interacting treatments are applied, this issue is especially pressing.

Some of the practical limitations of past RCTs relating to social interventions have long been recognized even by doyens of the method and of meta-analyses of findings from systematic reviews prioritizing RCTs, as shown in the following statement from 1993:

> The proper agenda for the next generation of treatment effectiveness research, for both primary and meta-analytic studies, is investigation into which treatment variants are most effective, the mediating causal processes through which they work, and the characteristics of recipients, providers, and settings that most influence their results. Such a research agenda is justified by a basic assumption that psychological treatment can be, and generally is, effective, so the questions of interest are not whether it works but how it can be made to work better. (Lipsey and Wilson, 1993: 1201)[7]

Lipsey and Wilson's comments come at the end of a long review of studies that examine the effects of psychological treatment. They find some support for most treatments, but unequivocal support for few. This is typical of most interventions, which produce a mix of winners, losers and those unaffected. Even with net positive or net negative findings across treated populations, some participants may have been affected but in the opposite way from the net impact. As Lipsey and Wilson suggest, progress can be achieved by refining understanding so that interventions are better designed and better targeted. This is important for PPP decision-makers who seldom start with a blank sheet and who are attempting to produce better states of affairs and behaviours than would otherwise occur (or cheaper achievement of similar states of affair or behaviours). It is that PPP refinement which requires research

evidence focused on 'treatment variations', 'causal processes' and the 'characteristics of recipients, providers and settings that most influence results'. There is a clear affinity between this and the emphasis that realists place on mechanisms and contexts. It is this affinity that lies behind the two Ms of EMMIE.

The first M in EMMIE refers to 'Mechanisms' (or 'Mediators'), the second 'M' to 'Moderators' (or contexts). 'Mechanisms' clearly allude to the causal forces stressed in realist evaluation and 'mediators' refer to the 'causal processes' (generally intervening variables) whose importance is emphasized by Lipsey and Wilson. Moderators refer to the variables sometimes used in statistical analysis to deal with effects variation in individual studies and across studies in meta-analysis, for example across study designs, recipients, providers or settings, while context is crucial to the realist's understanding of the contingent activation (or deactivation) of causal mechanisms, due to their dependency on specific attributes of recipients, providers or settings (on this see Cartwright and Hardie, 2012; Pawson and Tilley, 1997).

Some realists may object that associating mediators with mechanisms and moderators with context fails to recognize or acknowledge the gulf between the ontological, epistemological and methodological assumptions that lie behind black box RCTs (and their quasi-experimental counterparts) and those that inform realist evaluations. RCTs have tended to assume constant conjunction accounts of causality rather than generative ones. Moderators and mediators have comprised efforts to come to more refined constant conjunction causal conclusions. For realists, causality is generative and 'mechanisms' refer to generators that lie behind observable conjunctions. They depend on the activation of the causal potential from the intervention. That activation is contingent on sufficiently conducive conditions, which need to be specified. All that acknowledged, what matters for the PPP decision-maker is to understand that the production of intended and unintended outcomes is dependent on the possibility that the intervention will lead to those outcomes and the conditions for the actual production of those outcomes. Moreover, they need to understand this at a level of abstraction that goes from past specifics to present specifics. For this they need to grasp how the intervention produces its outcomes and what is needed for that outcome-production to occur. Whether this is framed as mediators and moderators or mechanisms and contexts does not matter much for practical decisions, however much it might vex methodologists. The practical importance of Mechanisms and Moderators (MM) is highlighted in Tilley's study of attempted replications of a successful burglary reduction programme, whose mechanisms and contexts were underspecified in the original study leading to substantial divergences in what was delivered in practice with consequent unsurprising variations in outcome produced (Tilley, 1996)

The 'I' of EMMIE refers to 'Implementation'. It is used here to refer to the conditions for putting the planned interventions into place as intended. Implementation is the focus of much process evaluation. Process evaluations are useful in understanding what is actually delivered and how this may come to depart (sometimes radically) from the policy, practice or programme as formally conceived. For some programmes it is, of course, difficult to separate the implementation theory from the intervention theory. A case in point is the British Crime Reduction Programme whose realist assessment concluded that it was rooted in a web of erroneous context-mechanism-output/outcome assumptions (Tilley, 2004). Absent understanding what is involved in implementing a programme successfully, PPPs will be unable to emulate past successes. Moreover, negative outcome results from a trial may lead to the premature withdrawal of an intervention programme, whose failure was a consequence of some aspect of

implementation weakness, which is liable to occur in particular in embryonic programmes involving long causal chains and/or multiple interventions (see Campbell, 1984). The PPP decision-maker needs evidence on implementation in order both to interpret failure and to appreciate what is involved in delivering a given intervention.

The 'E' of EMMIE refers to economic consequences. The rational PPP decision-maker needs to take account of these and economists have tried to monetize costs and benefits, even where these relate to apparently incommensurate outcomes, in order to create a common unit of account for comparative resource allocation calculations (for the variety of forms of economic evaluation see Manning et al, 2016). This has led some economists to be sympathetic to black box RCTs and efforts to emulate them as closely as possible where RCTs themselves are not practicable. Combined with conscientious and comprehensive costing, findings can feed into cost-effectiveness analyses (cost per unit of outcome or output) and estimations of effect sizes whose units might then be monetized and then compared to costs. The realists' recognition that outcomes are variable and unstable clearly creates a challenge for this, but the practical need for estimating expected monetized outcomes from different resource allocation decisions is difficult to deny.

Thus, in summary EMMIE refers to Effect (size), Mechanism (or mediator), Moderator (or context), Implementation, and Economic consequences. These capture the crucial empirical evidence needs for PPP decision-making. PPP decision-makers need to know what, in their circumstances with their target populations, can be expected to produce which outcome patterns at what costs and with what expected benefits. They need also to know what is involved in the successful delivery of the interventions that can produce given outcome patterns.

An EMMIE coding frame has been devised to rate and rank all systematic reviews of interventions attempting to reduce crime (Tompson et al, 2015). The results can be found at http://whatworks.college.police.uk/toolkit/Pages/Toolkit.aspx, which scores 37 reviews in terms of the quality of evidence and findings relating to Effect, Mechanism, Moderator, Implementation and Economic Cost. It has been sobering to find how thin the evidence base is. Further reviews are being conducted with an EMMIE focus. It is clear that primary studies have not been designed to meet EMMIE needs and in that sense, if the foregoing arguments are correct, they are thereby of rather limited use to PPP decision-makers. Primary evaluation studies would better meet PPP decision-maker needs if framed in EMMIE terms. This, in turn, would mean that commissioners of evaluations would need to frame their invitations to tender or requests for proposals in those terms.

The second half of this article assumes EMMIE but suggests that a black box clinical trial model of evaluation evidence, where RCTs are taken as the gold standard, may not the best starting point if we are interested in EMMIE usefulness in our work. The argument is presented here as a cockshy to try to provoke some fresh thinking.

EMMIE and engineering

Much evaluation has focused on health-related interventions. More particularly, recent discussions of evaluation standards have implicitly or explicitly used clinical trials as their source of a gold standard, albeit that that theoretical and empirical preparatory research that lies behind some clinical trials is often absent in other PPP domains. Medicine enjoys high status. The advances in medicine are obvious. It is unsurprising that there should be efforts to learn from

and emulate it. Although not originating in medicine, RCTs have become a taken-for-granted requirement for assessing and approving clinical treatments in medicine. Moreover, their limitations in health-related research have often formed the crux of methodological debates in evaluation.

This part of the article will argue, however, that engineering may provide a more instructive model for evaluation that aims to serve those PPP evidence needs of decision-makers captured by EMMIE. Engineers, however, do not write much about methodology (but see Petroski, 1982, 1996, 2008; Vincenti, 1990). They appear to be comfortable just doing it. And philosophers of science have likewise paid little attention to engineering (but see Scharff and Dusek, 2014 for a collection that shows some attention, but little to methodology). The reader is warned therefore that the following argument uses an understanding of engineering that is rather informal. It is derived in part from membership of the engineering faculty of a university and in part from readings in the history of aeronautics and in the sociology of science and engineering.

Many engineers promote some particular technology, gismo or material that they have developed. They are exponents of the law of the hammer: they have developed their counterpart to the hammer and aspire to hit everyone and everything with it. They are not EMMIE engineers. EMMIE engineers, instead, attempt to design, develop and create a physical, humanly constructed means for achieving an objective or an improved way of achieving that objective. When Popper (1957) referred to piecemeal social engineering he was talking about fashioning the social rather than the physical with the specific purpose of harm reduction.

Let us take aeronautical engineering as an example and reconstruct its methods. Orville and Wilbur Wright were, famously, the first to build a successful fixed wing airplane. They describe their methods (Wright, 1954), perhaps because they had no formal engineering training leading them to take for granted what they did. There has also been some effort to reconstruct their methodology (Johnson-Laird, 2005, 2006).

The Wrights' method differed from that used by those making earlier efforts at manned flight, whose approach seemed to be to have a bright idea and try it out.[8] Many perished as a consequence. They were fatal casualties of the tests of their hypotheses. The first to try and die was apparently King Bladud in 850BC. It was close to 2000 years later in 1903 that the Wrights succeeded. Repeat efforts in the meantime might have led classic trialists to conclude than manned fixed wing flight does not work![9] But the Wrights still tried and eventually succeeded, subsequently dying of natural causes. How come?

The Wrights began by reading all they could find on flight. They wrote to the Smithsonian Institute to gather together the material and they read it critically. They also read and built on the work of Sir George Cayley (1773–1857), a prolific engineer and inventor who developed a theory of flight, calculating what would be needed for a fixed wing plane to fly, without putting it to empirical test himself – he too died in his bed (Ackroyd, 2011). The Wright brothers worked out in detail what would be needed for flight according to the best theory they could find and they then subjected the crucial assumptions to empirical test. They even built a wind tunnel in the family bicycle manufacturing premises to test empirically the theoretically expected wing needs to achieve elevation. They developed a lightweight internal combustion engine that would produce sufficient power to deliver the output needed to drive the plane at a speed that relative to its weight and design would produce lift-off. They devised a design for propellers, based on analogous thinking related to (twisted) wings of birds (rather than propellers uses in ships), which would create the required drive. Each element of the design and its

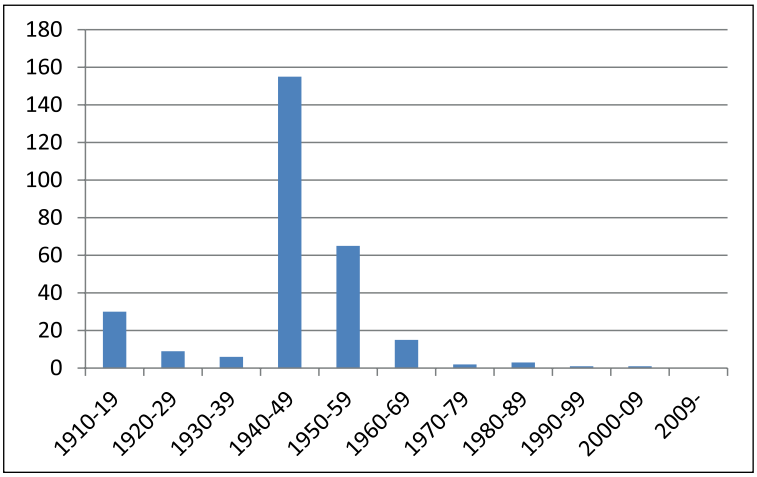

Figure 1. Trend in numbers of British test pilot fatalities: 1910.
Source: Calculated from data at http://www.thunder-and-lightnings.co.uk/memorial/index.php.

connection to other elements was worked through and tested separately. Some theory, for example, of the internal combustion engine, could more or less be taken for granted. All that which could not be taken for granted was articulated and tests undertaken before assembling the prototype which was again tested and tinkered with prior to the addition of a human, minimizing the risk of loss of life. The prototype test still mattered, of course. The plane might still have fallen from the sky and Orville Wright, who piloted it, might have perished. The first successful fixed wing flight required no RCT!

Following the Wrights' success the subsequent history of aeronautics has been one of attempted replication and then refinement after refinement, with a generally declining number of fatalities for test pilots. Figure 1 shows the clear trend in numbers killed.

Airplanes have become bigger, faster, safer and capable of greater distances between stops and thereby more cost-effective. This has not been achieved by RCTs or kindred trial designs. It has also involved trying whole new planes only after extensive testing of their components and near certain theoretical expectations that they will fly.

The core curriculum of trainee aeronautical engineers is instructive. One example is shown in Figure 2 taken from Teeside University.[10] It is clear from this that these engineers learn a lot of theory. Like the Wright brothers, modern aeronautical engineers also use a wide range of research methods including laboratory experiments, field trials, observations, simulations, thought experiments, historical comparisons etc. They do whatever is necessary to test and refine the working theories built into the airplanes being designed.

One general model for engineering methodology, drawn from Donald Campbell's ideas on 'evolutionary epistemology' (Campbell, 1974), has been referred to as 'selection-variation'. This begins with a hypothetical general solution to a problem, which is often rooted in theory but may also come from experience or intuition. Crucially what follows are tests of multiple variations in the translation of that hypothetical solution into practice, using whatever methods are suitable to the issue at hand. What best survives these tests then feeds into further iterations of the problem-solution and further variations that are then tested. Vincenti illustrates this process in the design and testing of profiles of airplane wings, where wind tunnels were used

Year 1	Year 2
Aerospace group design project	Aero engines and rocket science
Electrical principles	Aerospace group design and build project
Engineering design and CAD	Aircraft performance and stability
Engineering mathematics	Aircraft structures and materials
Engineering thermodynamics and heat transfer	Analytical techniques for engineers
Fluid mechanics	Avionics and aerospace systems
Professional skills for aeronautical engineers	Dynamic analysis
Properties of materials	Engineering management and leadership skills
Structural mechanics	

Figure 2. Core modules for year 1 and 2: Teeside BEng (Hons) Aerospace Engineering.

(Vincenti, 1990: 16–50; 241–50[11]). Vincenti stresses than selecting possible solutions for test-by-variation may involve 'hidden mental activities' that include searching 'past experience for similar situations to find knowledge that has proved useful', 'conceptual incorporation of whatever novel features come to mind as called for by new circumstances', and 'mental winnowing of the conceived variations to pick out those most likely to work'. Not everything, thus, is tested empirically and the choice of what to vary and what to test is far from random, although always blind to outcome. To Campbell the variation-selection mechanisms lie at the heart of progress in science, social programmes and technology. In some cases RCTs may comprise one of technique for testing, but it is not the only one.[12]

Of course airplanes, as with all forms of engineering and social programming, are not infallible. Many airplanes have fallen out of the sky and tragically it still happens. It is also the case that the modern aeronautical engineer has, perforce, to design to specifications that are adequate to multifaceted contextual conditions, which, if ignored, would lead to flight failures. Distances between airports, length of runways, risks of highjacking, emission targets, noise targets, mentally unstable pilots, extreme weather conditions, structural strains on materials, hostile government threats, etc. all comprise constraints on design that provide conditions for flights to deliver their intended outcomes (Petroski, 1996). Engineering designs of all kinds are also always lodged in wider social and physical systems whose functioning is crucial to the engineering outcome. These relate not only to the obvious processes of construction and control, but also to enabling or disabling settings, such as transport systems, public health, communications, maintenance, and access control (Petroski, 1996). Manifest failures lead to a form of diagnostic evaluation. What went wrong? Where was the theory built into the aircraft (or other engineering design) flawed?

Catastrophic failures of aircraft are by now rare and subject to detailed forensic examination. The crashes of the De Havilland Comets in Calcutta in May 1953, in Rome in January 1954 and Elba in April 1954 furnishes a useful example. Here was an airplane that had previously been reliable and safe, but crashes were now occurring. Where was the design flaw (Petroski, 1992)? Three crashes occurred before a common problem was identified. It is not easy to determine what caused a crash. As we saw with the Wright brothers, all airplanes are made of many crucial components rooted in many theories. Large jet liners such as the Comet incorporate much theory including that of materials and the conditions under which they will fail.

The Comet was built to tolerances that exceeded the expected strains that might lead materials to fail. Early mooted explanations for the initial crashes invoked extreme and abnormal weather conditions and pilot error. The third crash led to the view that there must be some hidden intrinsic weakness that had led the plane to fail repeatedly. A series of hypotheses were proposed that would explain what changed conditions (context) had unexpectedly caused (activated or deactivated causal mechanisms) to produce the catastrophic crashes (outcome pattern). In the end, the hypothesis that prevailed was that the cabin of the Comet had exploded (established by reconstructing how an Indian Anna – a type of coin – had left its impression on the tail plane) following a structural fault at the corner of one of the windows, which was produced by the repeated pressurization and depressurization of the cabin.

This pressurization–depressurization hypothesis was accepted following an experiment that simulated the relevant processes. The experiment involved filling and emptying the fuselage with water while flexing the wings with hydraulic jacks some 3000 times (Petroski, 1992: 178). This led to cracks round the window. Differences in air pressure would then lead to an explosion if this occurred as the plane was actually flying. While the plane had been designed comfortably to withstand expected pressures the consequences of repeated pressurization and depressurization had not been anticipated. Once recognized the failures could, of course be remedied. Comets flew again and an important lesson was learned for aircraft designers.

A similar exercise in forensic assessment of failure followed the explosion of the Challenger space shuttle on the morning of 28 January 1986 after its launch from Cape Canaveral (Collins and Pinch, 1998). Here it turned out that rubber O rings in joints separating solid rocket boosters were responsible. Unusually cold weather at the time of the launch had rendered these O rings brittle, leading them to crumble. Burning gases could then escape and this precipitated the disaster. On this occasion much recrimination followed the tragedy, with some suggestion that the problem could have been anticipated and remedial action taken. The O rings that were used had been tested with water in ways akin to those used in the Comet to test the pressurization/depressurization hypothesis, but these did not simulate the extreme cold on 28 January 1986. These and other tests were not, however, deemed adequate by some engineers at the time. Nevertheless the rings were used. Here the social context – pressure to launch – may have trumped precaution among key decision-makers leading to a failure to remedy a recognized weak point in the design.

What emerges from Petroski's writings on engineering design and engineering failure (Petroski, 1992, 1996, 2006) is that there are always realist (context-mechanism) theories, assumptions, trade-offs, compromises, uncertainties and economic factors at work in the design, development and delivery of products. He also presents engineering as a cumulative enterprise where past failures feed into improved future designs. In aeronautical engineering, for example, this is hard to miss. He emphasizes the imperfections of initial designs and the need for practical tinkering to iron out implementation problems: he cites the example of the Boeing 747, where 1000 lb of shims were needed to get bits of the fuselage to fit with one another - a problem now better dealt with computer aided design (Petroski, 1996).[13] Finally, it is clear that engineering is highly pragmatic.[14] The general lack of methodological self-awareness is matched by a concern not with the truth or compatibility of high-level explanatory theory but the practical adequacy of theories in use, whatever the origins of the theories.

Most engineering is EMMIE compliant, at least implicitly. There is a primary concern with E (effectiveness): Did the airplane fly/not fly? There is a concern with MM (mechanisms/mediators and moderators/contexts): What made the plane fly/crash in the prevailing conditions? What are the leading Context-Mechanisms-Outcome (CMO) hypotheses to be

tested? There is a concern with I (implementation): Was what was designed delivered and delivered consistently in practice? There is a concern with E (economy). Was the design delivered within budget and was the product more economical than its predecessor or next best alternative?

Moreover, engineering projects are comparable to PPPs. They comprise proposed new solutions or improved solutions to human problems (behaviours such as movement of people and products or states of affairs such as being too hot or cold or wet or dry). They are also fallible. However, notwithstanding failures there has been cumulative development in engineering expertise, which is less evident in PPPs. This may have partly to do with the stuff of the social as against the physical stuff of engineering. Human intentionality, social meanings, long and concatenating chains of complexity and feedback from human action, and apparently greater variability and changeability in the social world and so on may mark the social from the physical and inhibit the progressive understanding and control achieved in engineering. While nature might shout 'No' in the physical world, diverse human beings may individually and collectively shout (or whisper) 'No' (or 'Yes') in the social world in the light of their sometimes fickle wants and understandings.

The difference between the social and the physical of engineering, however, are comparable to those between the social and biological and the progress in engineering is of a similar kind to the progress in medicine. What matters here is whether the evaluation methods of engineering provide a better model for evaluation than those used in clinical trials and whether the consequent success and failure findings from engineering could be more useful to PPP decision-makers than success and failure findings from the methods used in clinical trials.

Here are two examples of realist evaluation and review that accord to some degree with the engineering model presented here, albeit that they lack that direct concern with practical improvement found in engineering.

First, Wong et al. report a realist review of possible legislation to ban smoking in vehicles carrying children (Wong et al, 2011). They identify eight threats to legislation as a public health measure targeting passive smoking: 'problem misidentification, criminalization, compensating behavior, lack of public support, lobby group opposition, obfuscating the new regulations, low perceived threat of enforcement and insufficient enforcement resources.' (Wong et al, 2011: 4). They were able to report on published evidence of various sorts that they were able to uncover in relation to four of them: problem misidentification, lack of public support, lobby group opposition, and enforcement (combining low perceived threat of enforcement and insufficient enforcement resources). Wong et al thus abstract and articulate the theories that are built into the legislation and the achievement of its aims and then cast around for evidence that would speak to those theories. This is comparable to the efforts made by the Wright brothers to identify and test the theories they were able to prior to their first successful flight in 1903, albeit that the Wrights also undertook some experiments of their own where they found existing evidence absent or inadequate. They were also able to demonstrate in outcome terms the success of their preparatory work.

Second, in an article already referred to Tilley (2004) abstracted and articulated the theories that were built into the failed British Crime Reduction Programme (CRP) in a realist forensic examination of its failure. He contrasted the 'supposed to do' theory with the ways in which the programme had actually played out, drawing out findings to a level of abstraction that speaks to kindred large-scale programmes and which explains how they often fail.

Neither of these realist evaluation studies articulated all the CMO theories that were built into the policy and programme examined. Many of the theories could properly be taken for granted. Instead, the studies abstracted that theory where weak points seemed most likely and where those weaknesses could jeopardize a successful outcome. In engineering and PPP decision-making, spotting the weakest points in the underlying CMO theory is clearly going to make the biggest contribution to improvement, rather than corroborating the rightfully taken for granted. In forensic studies, however, following failure, disentangling the specific unanticipated failure from the rightfully taken-for-granted can be difficult, as in the case of the crashing Comets.

It is sobering to contrast the CMO theory-focused and theory-testing evaluation of the Comet crashes with the theory-specification and testing indifference of the meta-analysis of Scared Straight (Petrosino et al, 2002), which found that criminality seemed to be produced rather than reduced. The systematic review included a) no analysis of points where the scared straight theory might be wanting; b) no proposals for or tests of alternative theory that might explain the negative effects; and c) no findings identifying the sub-groups where mechanisms were activated to produce negative effects. Little is learned except that programmes like Scared Straight had produced some net negative effects where RCTs had been undertaken.

The Scared Straight meta-analysis can be contrasted with an article that begins with an RCT, as do the evaluations collected together in the Scared Straight review, but which tries to identify and diagnose long-term (23-year) subgroup negative side-effects by focusing on the mechanisms generating them in the social contexts of affected subgroup members. There are tacit CMO conjectures. Sherman and Harris (2015) identify higher long-term death rates among employed black victims of domestic violence, as compared to their non-black or unemployed counterparts where the perpetrator was arrested rather than simply warned for misdemeanor domestic violence.[15] The conjectured mechanisms relates to post-traumatic stress (PTSS) occasioned by the arrest. A brief quote from Sherman and Harris's article gives a flavour of their proposed explanation:

> The paradox (in the relationship between race, employment, and victims' vulnerability to PTSS mediating their risk of death in response to arrest) would be present if there was greater resilience (and less PTSS) among black victims who were unemployed than among those who were employed, and if unemployed black victims also showed greater resilience than white victims whether they were employed or not. That paradox may relate to the extent to which employed African-American women, in particular, are more vulnerable to stress stemming from a threat to their livelihood or a change in their status and identity resulting from the threat of loss of employment, in comparison to unemployed women. The latter not only had no job to lose; many could also predictably depend on welfare benefits (such as aid for dependent children) for their financial support. Only a minority of women in Milwaukee's concentrated poverty neighborhoods were employed 'strivers,' pursuing what Anderson (1999) calls a 'decent' code of conduct. They could reasonably have feared a far greater loss of status and respectability from their partner being arrested than most of the unemployed women, who had a different 'code' or self-defined identity unrelated to employment.

Sherman and Harris acknowledge that they lack a direct test for their conjectures, but the following accord well with the engineering approach outlined in the latter part of this article:

(1) the identification of a negative side effect;
(2) the attempted diagnosis of it; and
(3) recognition that the diagnosis needs empirical test.

The implication would not necessarily be to abandon arrest as a response to misdemeanor domestic violence, but rather to tinker with it to try to maintain outcome benefits while avoiding negative side-effects.

Conclusions

This article has proposed the following six main arguments.

First, evaluations should focus on meeting the evidence needs of policy, programme and practice decision-makers, even if that means crossing boundaries between what have often been seen as incommensurable methodologies.[16]

Second, EMMIE (Effect, Mechanism, Moderator/context, Implementation and Economy) comprise five major considerations for policy, programme and practice decision-makers in relation to which empirical evidence can play a crucial role. Hence, to be most informative for decision-makers those conducting evaluations and systematic reviews are advised to collect evidence relating to all of them. This comprises a pragmatic approach wherein the realist methods most relevant to contexts, mechanisms and outcome pattern configurations are married to traditional PPP experimental methods used in most effect and economy studies and to qualitative methods that are often used to investigate implementation.

Third, evaluation might usefully model itself on engineering with its pragmatic focus on using any and all methods that help test crucial theories built into designs, especially those that are most likely to fail and lead to critical weaknesses in outcomes for some PPP participants. The choice of method is a matter of its fitness for the purpose of theory testing. And theory adequacy is likewise a pragmatic matter of fitness for purpose in delivering wanted outcomes in a cost-effective way. Engineering theories take the form of CMOs. Those of interest in practice are those most likely to fail. Moreover engineering is chronically concerned with overall success (Effect) and with achieving reliability (Implementation) and with achieving results within the resource limits available (Economy).

Fourth, the received wisdom suggesting that clinical trials/RCTs provide a gold standard for all other evaluations should be rejected. This is different from saying that there is no value in clinical trials. They have their place. A simple RCT can show only that a (well-targeted) measure can have an effect or that the measure does not always have its intended effect (although see Fletcher et al. and Hawkins, this issue, for a discussion of ways in which RCTs may be framed within realist evaluation). In health and medical research the RCT is one method among many that furnish evidence in favour of or against decisions to recommend a given treatment to a given individual with a given problem. Clinical trials involving RCTs are normally undertaken once much other research, using quite other methods, has been conducted to establish that there are reasonable grounds for hoping that a given treatment will lead to benefits for a given population. Results can be used to inform questions about variations by subgroup and context. In engineering RCTs are rare. Evidence is mostly collected in other ways. These appear to have been important in leading to improvements in engineering in terms of cost, reliability, efficiency and adaptability to new settings.

Fifth, the social has *sui generis* properties that suggest that neither clinical trials nor engineering methods can be taken off the shelf and applied with confidence that the findings will match the robustness expected from those methods when applied in medicine or technology. Issues of intentionality, reflexivity and complexity conspire to make the social a more fragile setting for interventions, albeit that there may be levels of abstraction at which robust, middle-range and useable CMOCs can be identified.

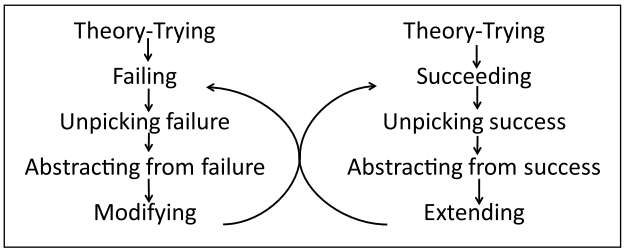

Figure 3. Processes of cumulation.

Sixth, and following Feyerabend, in terms of method 'anything goes', provided that it takes forward our understanding of what works for whom in what circumstances. Pragmatic considerations trump methodological purity.

Towards cumulation

The Holy Grail for applied as well as pure science is cumulation. Over the medium-to-long term this is obvious in engineering and health. It is less clear in social programmes. This may partly be to do with the distinctiveness of the social mentioned in my fifth point above. It may also have to do with the failure of research and evaluation to follow the kind of sequencing found in engineering and clinical medicine where successive studies seem to build on and be orientated to earlier ones rather than undertaken from scratch by generic evaluation journeymen. Figure 3 tries to represent cumulation as it has been achieved in clinical medicine and in engineering. It does not describe what occurs with PPPs. Perhaps it would be worth trying as a means of identifying and refining useable and portable CMOCs.

The left side of Figure 3 shows when an intervention fails in some respect (think of the crashing Comets). The failure is unpicked (think of the forensic examination of the Comet crashes and the tests of the hypotheses to explain the crashes). Abstraction from the failure occurs (think of the structural strains created by repeated pressurization/depressurization). Designs are modified to remedy the crucial weaknesses (think of the redesigned windows for the Comet). The right side of Figure 3 shows what happens where successes are achieved (think of the Wright brothers). Crucial features of success are identified and built into improved future designs (think the subsequent history of aeronautical engineering). Of course the two sides feed into one another as successes lead to extensions, which sometimes fail (the right to left feedback) and as failures lead to adaptations which sometimes succeed (the left to right feedback). The double feedback loops create increases in engineering achievement. They depend less on discrete tries of bright ideas and more on problem-solving and research development where what happens now builds on what happened yesterday, to extend success and to identify and remedy failures. The process depends on close working relationships between decision-makers and applied scientists orientated to helping decision-makers avoid errors and achieve successes. It is, one imagines, what Popper meant by piecemeal social engineering and Campbell by the experimental society (Campbell and Rosso, 1999). It is this model that promises most for cumulation in developing PPPs.

The implications of the foregoing discussion for the 'What Works' movement may be quite profound. They include:

- moving away from pass/fail verdicts or simple net impact estimates;
- concentrating heavily on identifying and diagnosing failures of and also within PPPs (even when the net effects appear positive);
- using diverse methods to test embedded programme theories and assumptions for their weakest points before, during and after PPPs are put in place;
- being explicit about causal mechanisms in successes and failures and the contexts needed for the activation and deactivation of those causal mechanisms;
- extensive tinkering with PPPs in their targeting and delivery-on-the-ground once the core ideas have been worked out and even during their implementation;
- an explicit agenda for adaptation and cumulation in PPP understanding to improve outcomes; and
- a continuing concern with best estimates of bottom-line economic effectiveness.

It will be difficult to overcome the 'what works' slogan itself even though it is freighted with misleading assumptions and expectations that cannot be met. However, evaluators and PPP decision-makers need to know better.

Acknowledgements

The original ideas relating to EMMIE were forged with my colleagues at UCL, Kate Bowers and Shane Johnson, with further work by Jyoti Belur and Lisa Tompson. Graham Farrell, Justin Jagosh, Gloria Laycock and Gill Westhorp read and commented on earlier drafts of the article. Referees of the submitted version made some excellent critical points, which rightly forced me to rethink several of the arguments. The article has thereby become more sober and less provocative, but I hope a little more plausible. As ever, all errors are mine alone. I am grateful for those corrections readers and referees have led me to make.

Funding

This article emerged from the Commissioned Partnership Programme: the What Works Centre for Crime Reduction, which has been funded by the Economic and Social Research Council (ESRC, grant ES/L007223/1) and the College of Policing, although the views expressed neither necessarily represent those of all (or indeed any other) members of the large consortium which has been involved in this research nor, indeed, those of the College of Policing.

Notes

1. Of course matters are often a little more complex, for example with no treatment, placebo, and conventional treatment often provided for one or more control group. The underlying logic, however, remains the same.
2. For one influential example of an evidence hierarchy, the Maryland Scale, see Sherman (1997). For a discussion of standards and a review of some others, see Nutley et al. (2013).
3. Systematic searches to find all studies that have the required methodological attributes are needed to avoid biases that may creep in if only those most readily available are drawn on. In particular a 'publication bias' towards intervention successes may occur if only those that are published are included.
4. The question of whether RCTs and quasi-experiments can be conducted in realist terms is touched on but the detailed argument lies beyond the scope of this article. Other contributions to this special issue discuss in different terms the possibility of making use of RCTs in the conduct of realist evaluations; see also Bonell et al. (2012, 2013) and Marchal et al. (2013).
5. Rothwell rightly emphasizes the risks of post hoc subgroup analysis. This may take the form of fishing expeditions. If sufficient variables are tried with large samples statistically significant subgroup variations will be expected as a matter of chance. This is not to say that trials cannot be

designed in advance to test for expected, theoretically informed subgroup variations or that apparently significant findings from fishing expeditions should not lead to further research to determine if there is, indeed, a causal relationship.

6. This heterogeneity problem continues in the meta-analyses of systematic reviews. A recent paper on trials has noted two ideal types, one of which is 'explanatory' and the other of which is 'pragmatic', although the authors note that these stand at opposite ends of a continuum (Loudon et al, 2015). Pragmatic trials attempt to assess the 'real world' effectiveness of an intervention to inform decision-makers about its adoption. Explanatory trials use idealized settings to 'give the initiative under evaluation its best chance to demonstrate a beneficial impact' (Loudon et al, 2015: 1). The authors note that few trials fit either ideal perfectly, and they have developed ('with the help of' 80 international trialists, clinicians and policymakers') a tool to place trials at one of five points along the continuum. The distinctions they make and the problems in making them highlight the difficulty a) in determining what can be learnt from any given trial and b) in synthesizing a set of trials that sit at different points (or maybe mostly towards one end) in the continuum Loudon at al usefully identify.
7. This statement could have been made by Pawson and Tilley and was made after their first sketch of realist evaluation, which appeared in quite an obscure edited collection (Pawson and Tilley, 1992). However, there is no reason to believe Lipsey and Wilson were drawing on Pawson and Tilley, to whom they make no reference.
8. This may still be the case with many social interventions, where neither thought through theory nor relevant evidence of past effectiveness seem to lie behind them. In King and Crewe's (2013) catalogue of catastrophes, the millennium dome and poll tax are examples. Other cases show well-developed theory that evaluators can help formalize and test (see Tilley, 2015).
9. For a (realist) discussion of rational decisions not to accept falsifying evidence, see Koslowski (1996), and for the need for fiddling with apparatus, contact with others who had succeeded and the tacit knowledge thereby gained in translating the possible into something that works, see Collins (1985).
10. A brief look at the curricula at other universities suggests that this is fairly typical. Some do include modules on experimental methods, but these relate mainly to the practicalities of testing and measurement.
11. Although not framed in these terms, this approach is exemplified in Syed (2015). Campbell, Syed and Vincenti all hark back to Popper as the inspiration for their writings on cumulative developments in science and engineering.
12. Campbell (1974: 435) interestingly notes that scientific progress often speeds up where discoveries such as the 'barometer, microscope, telescope, galvanometer, cloud chamber, and chromatograph' have made testing hypotheses easier. This speaks to the importance of taking advantage of diverse and emerging methods of testing PPP theories.
13. Gawande (2010) has made a similar point about engineering and the construction of skyscrapers where, notwithstanding initial designs, in the translation of blueprints on the ground tinkering is needed to overcome unanticipated problems. Gawande's own account of the developments in surgery checklists likewise shows the need for tinkering in a social programme.
14. Pragmatic is used here to mean, 'for practical purposes'. It is not used to refer to philosophical pragmatism. See also note 16.
15. Some readers may suspect homicide. This was not the case.
16. For a more extensive (and realist-related) discussion of the focus of research aiming to inform improvements in policy, programme and practice, focusing on crime but with general implications, see Tilley (2016). Wider debates on use and usability are, however, complex and lie beyond the remit of that chapter or this article.

References

Ackroyd J (2011) Sir George Cayley: The invention of the aeroplane near Scarborough at the time of Trafalgar. *Journal of Aeronautical History* Paper No 2011/6: 130–81.

Anderson E (1999) *Code of the Street: Decency, Violence, and the Moral Life of the Inner City*. New York: W.W. Norton.

Bonell C, Fletcher A, Morton M, et al. (2012) Realist randomized controlled trials: A new approach to evaluating complex public health interventions. *Social Science and Medicine* 75: 2290–306.
Bonell C, Fletcher A, Morton M, et al. (2013) Methods don't make assumptions, researchers do: A response to Marchal et al. *Social Science and Medicine* 94: 81–2.
Brynner R and Stephens T (2001) *Dark Remedy*. New York: Basic Books.
Campbell D (1974) Evolutionary epistemology. In: Schilpp P (ed.) *The Philosophy of Karl Popper*, Vol. 1. La Salle, IL: Open Court, 413–63.
Campbell D and Russo M Jean (1999) *Social Experimentation*. London: Sage.
Campbell D (1984) Can we be scientific in applied social science? In: Conner R, Altman D and Jackson C (eds) *Evaluation Studies Review Annual*, Vol. 9. Beverley Hills, CA: SAGE, 26–48.
Cartwright N (2007) Are RCTs the gold standard? *BioSocieties* 2: 11–20.
Cartwright N and Hardie J (2012) *Evidence-Based Policy*. Oxford: Oxford University Press.
College of Policing and ESRC (2013) *Commissioned Partnership Programme in Support of the What Works Centre for Crime Reduction*, Unpublished Call for Proposals.
Collins H (1985) *Changing Order*. London: SAGE.
Collins H and Pinch T (1998) *The Golem at Large*. Cambridge: Cambridge University Press.
Davidoff F (2009) Heterogeneity is not always noise: Lessons from improvement. *JAMA* 302(23): 2580–6.
Feyerabend P (1975) *Against Method: Outline of an Anarchist Theory of Knowledge.* London: New Left Books.
Gawande A (2010) *The Checklist Manifesto*. London: Profile Books.
Greenhalgh T, Howick J and Maskrey N (2014) Evidence based medicine: A movement in crisis. *BMJ*: 348: g3725.
Halpern D (n.d.) What works: Creating an evidence-based approach. what-works-david-halpern.ppt. Cabinet Office. Available at: http://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0CB0QFjAA&url=http%3A%2F%2Fwww.solace.org.uk%2Fevents%2Fwhat–works–david–halpern.ppt&ei=BBhoVaiEEejD7ga164OwDw&usg=AFQjCNFC3–GAmZ0zSDAL0udKlSFLYEQDKA&bvm=bv.93990622,d.ZGU (accessed 29 May 2015).
Her Majesty's Government (2012) *The Civil Service Reform Plan*. London: Cabinet Office.
Her Majesty's Government (2013) *What Works: Evidence Centres for Social Policy*. London: Cabinet Office.
Johnson S, Tilley N and Bowers K (2015) Introducing EMMIE: An evidence rating scale to encourage mixed-method crime prevention reviews. *Journal of Experimental Criminology* 11(3): 459–73.
Johnson-Laird P (2005) Flying bicycles: How the Wright brothers invented the airplane. *Mind and Society* 4: 27–48.
Johnson-Laird P (2006) *How We Reason*. Oxford: Oxford University Press.
Kent D, Rothwell P, Ionnidis J, et al. (2010) Assessing and reporting heterogeneity in treatment effects in clinical trials: A proposal. *Trials* 11: 85.
King A and Crewe I (2013) *Blunders of our Governments*. London: Oneworld.
Koslowski B (1996) *Theory and Evidence.* Cambridge, MA: The MIT Press.
Lipsey M and Wilson D (1993) The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist* 48(12): 1181–209.
Loudon K, Treweek S, Sullivan F, et al. (2015) The PRECIS-2 tool: Designing trials that are fit for purpose. *BMJ* 350: h2157.
McBride W (1961) Thalidomide and congenital abnormalities. *The Lancet*, December: 1358.
Manning M, Johnson S, Tilley N, et al. (2016) *A Guide to Economic Analysis and Efficiency*. Basingstoke: Palgrave.
Marchal B, Westhorp G, Wong G, et al. (2013) Realist RCTs of complex interventions – An oxymoron. *Social Science and Medicine* 94: 124–8.
Mukherjee S (2011) *The Emperor of all Maladies*. London: Fourth Estate.
Nutley S, Powell A and Davies H (2013) *What Counts as Good Evidence?* St Andrews: University of St Andrews Research Unit for Research Utilisation.

Pawson R and Tilley N (1992) Re-evaluation: Rethinking research on corrections and crime. In: Duguid S (ed.) *Yearbook of Correctional Education*. Burnaby: Institute of Humanities, Simon Fraser University, 19–49.
Pawson R and Tilley N (1997) *Realistic Evaluation*. London: SAGE.
Peto R, Collins R and Gray R (1995) Large-scale randomized evidence: Large, simple trials and overviews of trials. *Journal of Clinical Epidemiology* 48(1): 23–40.
Petrosino A, Turpin-Petrosino C and Buehler J (2002) 'Scared straight' and other juvenile awareness programs for preventing juvenile delinquency. In: *The Campbell Collaboration Reviews of Intervention and Policy Evaluations* (C2-RIPE). Philadelphia, PA: Campbell Collaboration.
Petroski H (1982) *To Engineer is Human*. New York: Random House.
Petroski H (1992) *The Evolution of Useful Things*. New York: Random House.
Petroski H (1996) *Invention by Design*. Cambridge, MA: Harvard University Press.
Petroski H (2008) *Success through Failure: The Paradox of Design*. Princeton, NJ: Princeton University Press.
Popper K (1957) *Poverty of Historicism*. London: Routledge.
Rothwell P (2005a) External validity of randomized controlled trials: 'To whom do the results of this trial apply?' *Lancet* 365: 82–93.
Rothwell P (2005b) Subgroup analysis in randomized controlled trials: Importance, indications, and interpretation. *Lancet* 365: 176–86.
Scharff R and Dusek V (eds) (2014) *Philosophy of Technology: The Technological Condition, an Anthology*. Chichester: Wiley, Blackwell.
Semmelweiss I (1983 [1860]) *Etiology, Concept and Prophylaxis of Childbed Fever*. Madison, WI: University of Wisconsin Press.
Sherman L (1997) Thinking about crime prevention. In: Sherman L, Gottfredson D, Mackenzie D, et al. (eds) *Preventing Crime: What Works, What doesn't and What's Promising?* Washington, DC: Office of Justice Programs, Chapter 2.
Sherman L and Harris H (2015) Increased death rates of domestic violence victims from arresting vs. warning suspects in the Milwaukee Domestic Violence Experiment (MilDVE). *Journal of Experimental Criminology* 11: 1–20.
Syed M (2015) *Black Box Thinking*. London: John Murray.
Tilley N (1996) Demonstration, exemplification, duplication and replication in evaluation research. *Evaluation: The International Journal of Theory, Research and Practice* 2(1): 35–50.
Tilley N (2004) Applying theory-driven evaluation to the British Crime Reduction Programme: The theories of the programme and of its evaluations. *Criminal Justice* 4: 255–76.
Tilley N (2015) There is nothing so practical as a good theory: Teacher-learner relationships in applied research for policing. In: Cockbain E and Knutsson J (eds) *Applied Police Research: Challenges and Opportunities*. London: Routledge, 141–52.
Tilley N (2016) Middle range radical realism for crime prevention. In: Matthews R (ed.) *What is to Be Done About Crime and Punishment?* Basingstoke: Palgrave, 89–122.
Tompson L, Bowers K, Johnson S, et al. (2015) *EMMIE Evidence Appraisal Tool*. London: UCL Discovery. Available at: http://discovery.ucl.ac.uk/1462093/ (accessed 18 April 2016).
Vincenti W (1990) *What Engineers Know and How They Know It*. Baltimore, MD: The Johns Hopkins University Press.
Wong G, Pawson R and Owen L (2011) Policy guidance on threats to legislative interventions in public health: A realist synthesis. *BMC Public Health* 11: 222.
Wright O (1954) *How We Invented the Airplane*. New York: Dover Publications.

Nick Tilley is a Professor in the Jill Dando Institute of Crime Science, UCL. His long-term interests lie in theoretically informed applied social science. Most of his substantive work has focused on crime prevention and on policing. He co-authored *Realistic Evaluation* with Ray Pawson.