The Health Foundation
Inspiring Improvement

Perspectives on context

The problem of context in quality improvement

Professor Mary Dixon-Woods

About the authors

Mary Dixon-Woods is Professor of Medical Sociology and Wellcome Trust Senior Investigator in the SAPPHIRE group, Department of Health Sciences, University of Leicester, UK and Deputy Editor-in-Chief of *BMJ Quality and Safety*. She leads a large programme of research focused on patient safety and healthcare improvement, healthcare ethics, and methodological innovation in studying healthcare.

Contents

The problem of context in quality improvement

1. Introduction

Though (formal) quality improvement in healthcare has only a brief history, it is history littered with examples of showpiece programmes that do not consistently manage to export their success once transplanted beyond the home soil of early iterations,[1] or that demonstrate startling variability in their impact in apparently similar settings. Quality improvement (QI) collaboratives – involving multidisciplinary teams working across departments or organisations to address quality issues – are, despite their popularity, a good example of both of these effects.[2] It has been estimated that only 30% of organisations involved in collaboratives may achieve 'significant improvements' and that another 30% may drop out before the end.[3] One of the explanations most often advanced to explain the differential impact of QI efforts is that of context.[4,5] Jonathan Lomas[6] goes as far as suggesting that the 'overriding influence of context' may go a long way towards explaining why there remains no clear advice on how to go about improving quality in healthcare. But a review of strategies for improving quality and safety in healthcare has deplored the absence of attention to context and implementation factors.[7]

The term 'context' has its etymological roots in the Latin *contextus*, meaning 'joining together'. Understanding what happens when a particular QI intervention is joined together with a team, organisation, or health system, through multiple interacting contextual layers, is a challenge both for science and for practice and policy. Though the need for serious attention to questions of context has become increasingly well recognised across a range of disciplines, including politics,[8,9] only latterly has the context sensitivity of many QI initiatives in healthcare become properly recognised.[10,11] The challenge now is twofold: how to study interactions between contexts and interventions to develop a more credible science of quality improvement, and how to deal with contextual effects in implementing quality improvement interventions. But how we should structure thinking about context remains a stubborn puzzle.

In this paper, I want to make some proposals that may be helpful in moving the field forward. I will suggest that no account of context can be decoupled from a broader understanding of causation, and that our view of causation must include both assessment of whether inputs and outputs of interventions are correlated, **and why** such a correlation occurs. I will suggest that QI studies have much to learn from the clinical sciences, but I will also argue that current clinical science methodologies are not enough to gain a proper understanding of QI. Along the way, I will take issue with various arguments put forward by the advocates for 'realist evaluation', suggesting that some of these arguments offer little that is distinctive and that others are mistaken. In particular, I will propose that while the need to construct explanatory accounts of interventions (or identify the causal mechanisms, as realists have it) is essential, abandoning a correlational approach in the process is reckless and misguided. I will suggest that a focus on **practical wisdom** and a blend of methods from the clinical and social sciences is likely to offer the best way forward.

2. Context and causality

Consideration of context goes hand in hand with the problem of establishing causality. In clinical epidemiology, the standard approach to determining causality is based on statistical reasoning. Determining whether an intervention is responsible for an observed effect is worked out through **correlational** logic. Studies using this approach are concerned with assessing whether, **on average**, the independent variable (the intervention) has made a measurable difference to the dependent variables (the outcomes). This can be done by following those that received the intervention over time, but it can be hard to rule out the possibility that any improvement detected was not really due to the intervention but to some other cause or causes. These possible influences on the outcomes other than the intervention include 'contextual factors'. Attempts to isolate the effect of an intervention therefore seek to consider systematically (for example, through regression modelling techniques) the extent to which 'confounding' factors might be responsible for any observed change. The ease with which the effects of these confounders can be detected and assessed is greatly improved by manipulating the inputs – for example, by having one group receive the intervention and another act as the control (a controlled study). Randomised controlled trials (RCTs) are therefore seen as the most powerful design for establishing a causal relationship.

Realist evaluation

A series of challenges to this standard approach has been offered in recent years, most prominently by those working within the 'realist evaluation' paradigm. This approach seeks to abandon the correlational view of causality and substitute it with one focused on identifying and assessing the **mechanisms** that explain configurations of contexts-mechanisms-outcomes. The social scientists Ray Pawson and Nick Tilley are most strongly associated with this approach, which is presented in manifesto form in their 1997 book *Realistic Evaluation*.[12] Because they have become well known (and to some extent influential) within the health sciences, their work provides a useful point in which to anchor discussion.

Pawson and Tilley[12] argue that experimental approaches are 'black boxes' that only describe outcomes, not explanations of why programmes work or fail. Such approaches are argued to neglect the significance of context. Pawson and Tilley condemn what they see as the **successionist** logic underlying the RCT model, urging instead adoption of a **generative** theory of causation. Successionist approaches, they argue, determine causality on the basis of co-variation, and assume that the cause of change is external and will consistently produce the same effect. Generative theories – by contrast, they argue – accept that causal relationships may be linked to an external intervention but assume that the impact of the intervention also depends on internal features or characteristics of the context in which the intervention is introduced. A key assumption of realist evaluation is that programmes have differential effects because the mechanisms responsible may not be activated in all contexts.

Rejecting the tendency to treat contextual variables as 'confounding', Pawson and Tilley propose that the contexts within which causal mechanisms operate should be the focus for understanding. They seek to identify the different ways in which contexts, mechanisms, and outcomes can be 'configured', and propose that theory can be tested and developed though a process of comparison of 'families of configurations'. Realist evaluations typically reject the idea that programme 'success' can be determined through the performance measures characteristic of correlational evaluation, as a recent study in this tradition again demonstrates.[13]

Pawson and Tilley's exhortation to identify 'what works for whom in what circumstances' is certainly beguiling rhetoric. It probably explains the appeal of realist evaluation to those frustrated by the zealotry associated with some of the evidence-based medicine movement, including the insistence that the only legitimate source of knowledge is that which is countable or measurable. Pawson and Tilley are right to emphasise the need to theorise about the links between interventions and outcomes, and right about the need to attend to context. But while they are asking the right questions and giving some of the right answers, they are very far from unique in this in either the social or the clinical sciences; nor, in the end, do they offer a **methodological** solution for studying context and causation.

Theory building in the clinical sciences

It is a mistake, as Pawson and Tilley do, to dismiss controlled studies in medical sciences as relying on a flawed 'successionist' logic. Much scientific progress in clinical medicine is achieved through careful theorisation about possible mechanisms that might bring about desired outcomes, and through iterative testing of these theories through a range of study designs, of which the controlled study is a key element. A new therapeutic agent, for example, is typically based on theory (or a set of theories) about disease processes and the likely action of the agent in targeting these – the mechanism. Thus, for instance, the action of temozolomide in the treatment of brain tumours is theorised to involve methylation of DNA and consequent death of tumour cells. The development of this theory of the mechanism underlying the observed outcomes proceeded not through mechanistic deductions, but through an iterative, creative, and sometimes messy process of discovery, abduction, and testing.[14]

Therapeutic agents, far from being black boxes, typically go through multiple sequences and feedback loops of theory testing and refinement aimed at understanding how the agent will be processed by the body, and at determining the likely outcomes of the drug – both intended and unwanted. This kind of research thus follows the route commended by realist evaluation. Pharmacokinetic studies will be one of the earliest in the sequence of challenges to which the theory is exposed. Such studies attempt to determine the processing of the drug once in the body, by looking at how it is metabolised, what systems it acts on, and what kinds of biochemical and other changes it produces. Pharmacokinetic studies are, in Pawson and Tilley's terms, classically **generative** in their orientation (given the internal characteristics of the human body, what is the likely destiny of this external agent?). Such studies often result in what appear to be promising agents being abandoned, or in adjustments being made to the formulation of the drug, because the data emerging from the study improves knowledge (theory) about the likely mechanism of action. It seems clear that, if converted into Pawson and Tilley's terms, the proper way of understanding the drug in the body is to see it as an example of context + mechanism.

Such an understanding comes even more plainly into view in the latest developments in clinical science, which are increasingly showing how variations at the molecular level in individual patients influence the fate of drugs in the body. Variations in the genotypes of both individuals and population groups are now known to have profound influences on responses to drugs in terms both of effectiveness and adverse effects. For example, cancers that appear histologically similar (on examination by microscope) may turn out to be very different at the molecular level.[15] Pawson and Tilley make a great deal of their claim that different mechanisms can produce the same outcomes, and suggest that different context-mechanism configurations may produce the same outcome or the same context-mechanism configurations may produce different outcomes. They are not wrong, but theirs is not a unique insight: it is one that has been accepted within medicine for decades. Clinical research is increasingly showing how the same phenotype (say, asthma) may have its origins in very different genotypes, while the same genotypes may produce very different phenotypes.

Arguing that 'causal outcomes follow from mechanisms acting in contexts', Pawson and Tilley propose that interventions provide a trigger for change only if the prevailing conditions can support change: 'programmes work (have successful "outcomes") only in so far as they introduce the appropriate ideas and opportunities ("mechanisms") to groups in the appropriate social and cultural conditions ("contexts")'. They further argue that:

> *'Context describes those features of the conditions in which programmes are introduced that are relevant to the programme mechanisms... For realism, it is axiomatic that certain conditions will be supportive to the programme theory and some will not. And that gives realist evaluation the crucial task of sorting the one from the other.'*[16]

Again, Pawson and Tilley are right, but there is nothing specific to realist evaluation about their programmatic claims or aims. Modern clinical science is preoccupied with exactly the same kinds of tasks. Genetic heterogeneity may mean that only some patients in a target population have diseases that are treatment sensitive.[17] For example, variations in the MGMT gene strongly influence positive response to temozolomide,[18] but the existence of such a mechanism

can be established only through sophisticated application of a range of methods, including controlled experiments. Contrary to Pawson's claim that 'It would be an absurdity in most medical trials to imagine that the patient transforms the treatment',[19] that is precisely what happens. Some interventions are effective in some patients; some are not. This is because of genetic – that is to say, contextual – variation. Bodies are not regarded in modern clinical science as passive objects, nor is explanation of mechanism written out. It is no exaggeration to say that modern clinical science is now as much concerned with what bodies do to drugs (the impact of context on intervention) as it is with what drugs do to bodies (the impact of interventions on specific contexts). Further, clinical trials only exist and advance because of continual efforts to improve the theoretical bases of postulated mechanisms. For example, to continue with the temozolomide illustration, work is currently underway to improve treatments by combining the drug with other agents that are theorised to increase its potency in killing tumour cells. There is thus little that is distinctive about many of the programmatic aspirations of realist evaluation. Its conceptualisation both of the significance of context and of the need to identify the conditions supportive of an intervention, as well as theoretically informed explications of causal mechanisms, seems so closely to mirror that of modern, molecularly based clinical science as to be indistinguishable. 'What works for whom in what circumstances' seems as good a description of the aims of molecularly oriented clinical science as any.

What can quality improvement learn from the clinical sciences?

The key point in the discussion thus far is not simply to argue that, in some of its fundamentals, realist evaluation has far more in common with clinical science than its proponents might think (or want to accept). Rather, I have three aims. First, comparing clinical science and realist evaluation allows us to dispose of the idea that controlled study and experimental designs are somehow fatal to efforts to investigate context and identify causal mechanisms. Embracing a correlational logic does not mean that context is somehow ignored or distorted, and that interest in characterising causal mechanisms evaporates. Pawson and Tilley argue that *'when we explain a regularity generatively, we are not coming up with variables or correlates which associate one with the other; rather we are trying to explain how the association itself comes about'*. Yet without a sound understanding of whether and how variables or correlates are associated, many attempts to construct explanation are doomed.

Without using quantitative modelling, for example, many structural-level influences relevant to theory building may remain obscured. Kieran Healy's work, for instance, shows that any effect of presumed consent laws on rates of organ donation can be explained by attention to the social organisation of transplant systems in countries that have implemented such laws.[20] High-yield countries such as Spain and Italy do not owe their success to different legal rules from those in opt-in countries, but to effective investment in system logistics and management: they have more staff dedicated to the procurement process, more training in getting consent from families, and improved coordination within the system. This conclusion about the contextual influences on organ donation, and the mechanisms implicated in donation, was reached by Healy following detailed quantitative analysis **and** highly sophisticated application of theory from institutional and economic sociology. In throwing out the correlational baby with the bathwater, realist evaluation risks failing to provide the kind of evidence needed to establish the effectiveness of interventions or to identify the institutional and structural aspects of context that are potentially open to remedy.

Second, claims about the supposed defects of controlled trials and experimental design are a distraction from the real work of improving the science underlying quality improvement in healthcare, including the problems of context. The problem lies not so much in fundamental defects of the clinical science/clinical trials approach to drug development, as in the unsophisticated application – or non-application – of many of its more useful principles **to implementation and study of QI interventions**. By the usual standards of clinical epidemiology, QI research is a field – ironically – beset by quality problems, including widespread use of study designs that limit the confidence with which change can reliably be attributed to the intervention,[21] use of poorly operationalised measures of both programme inputs and outcomes, poor quality of data collection, and reluctance to search for unintended consequences or determine the cost-effectiveness of interventions.

One striking feature of the development of QI interventions, for instance, is their tendency to neglect the equivalents of the laboratory and the pre-clinical and pharmacokinetic stages of drug development. QI interventions tend to move straight to implementation, bypassing the stages of characterising the intervention and exploring how it is that individuals, teams or organisations 'metabolise' the intervention, what likely contextual influences might neutralise or subvert the intervention, what the unwanted effects might be, or how any observed effects can be properly explained. Thus, for instance, QI interventions have been routinely imported from very different sectors – such as aviation – without adequate consideration of whether these sectors are more like different species (zebras being compared with lions). Improving the design and execution of studies of QI in order to provide more reliable evidence is a priority.

This links to the third reason for drawing attention to the evolving paradigm within the clinical sciences, which is to highlight the level and quality of methodological innovation now occurring to deal with the recognition of the complexity of gene–environment interactions. For example, it is clear that unrecognised molecular heterogeneity can reduce the power of randomised trials to detect therapies that may be beneficial for specific subgroups, and statistical techniques are under development to deal with this challenge. New and emerging techniques in genetic epidemiology[22] also represent a rich treasury of methods. With some adaptation, such approaches might be applied to modelling contextual variables relevant to QI efforts, and thus enhance the ability to make much better assessment of risk factors for the implementation of QI activities – including the kinds of factors likely to leave such activities incapable of delivering any benefits. A recent review by Shekelle et al[11] identified four salient areas of context influencing patient safety practices in healthcare organisations:

a. Structural organizational characteristics (such as size, location, financial status, existing quality and safety infrastructure).

b. External factors (such as regulatory requirements, the presence in the external environment of payments or penalties such as pay-for-performance or public reporting, national patient safety campaigns or collaboratives, or local sentinel patient safety events).

c. Patient safety culture (not to be confused with the larger organizational culture), teamwork, and leadership at the level of the unit.

d. Availability of implementation and management tools (such as staff education and training, presence of dedicated time for training, use of internal audit-and-feedback, presence of internal or external individuals responsible for the implementation, or degree of local tailoring of any intervention).[11]

It is likely that many (though not all) of these contextual variables are capable of being measured and then modelled. This kind of analysis can help in building insights into where efforts need to be targeted, and what preparations organisations need to make when introducing QI initiatives. However, as I shall emphasise later, quantitative models on their own will never be enough to ensure a full accounting for context both in the study and implementation of quality improvement efforts, nor a full explanation of how interventions lead to outcomes.

3. Why the clinical science approach is not enough

One of the real achievements of those working within the mechanismic paradigm (including realist evaluation) has been to refocus attention on the need for better understanding of **what it is that is causing the changes observed** rather than being content simply to determine a causal effect and the confounding variables that modify such effects. Yet studies of QI typically suffer from two major problems. First, they are often remarkably poor at describing exactly what the intervention comprises within reports, and often fail to characterise the intervention and its activities in such a way that it can easily be reproduced. Second, such studies are equally poor at describing the theoretical basis of their interventions (what is the means by which this intervention might reasonably be expected to achieve the hoped-for effects?).[7] Further, attempts to update theories in response to the findings of empirical studies based either on process evaluation or learning acquired during the running of the intervention remain rare in QI, so that theory evolution remains stunted. As Shojania and Grimshaw note:

'From the perspectives of clinical medicine and the research enterprise, we regard it as absurd to proceed directly from a patient's poorly

understood complaints to reaching for a bottle of pills simply because they are handy and resemble ones recommended anecdotally by a colleague. The decision to administer these pills without any understanding of their active ingredients or their mode of action would be completely unsupportable. Yet comparably unsupportable activities occur routinely in quality improvement (QI) research.'[23]

Cargo cult quality improvement

The failure to produce good quality accounts of what the intervention involved (what were the activities undertaken?) and the theory explaining how the intervention achieved its effects (what mechanisms were at work?) leads to a number of important problems for QI, including the problem that might be termed 'cargo cult quality improvement'. Cargo cult science was famously described by Richard Feynman in a 1974 Caltech commencement address:

'In the South Seas there is a Cargo Cult of people. During the war they saw airplanes land with lots of good materials, and they want the same thing to happen now. So they've arranged to make things like runways, to put fires along the runways, to make a wooden hut for a man to sit in, with two wooden pieces on his head like headphones and bars of bamboo sticking out like antennas – he's the controller – and they wait for airplanes to land. They're doing everything right. The form is perfect. It looks exactly the way it looked before. But it doesn't work. No airplanes land. So I call these things Cargo Cult Science, because they follow all the apparent precepts and forms of scientific investigation, but they're missing something essential, because the planes don't land.'[24]

When QI initiatives are implemented without proper understanding of what they involve and how they work, they similarly risk becoming pale and distorted imitations that succeed only in reproducing the superficial outer appearance, but not the mechanisms (or set of mechanisms) that produced the outcomes in the first instance.

There can be little doubt that cargo cult QI explains some of the variability in the outcomes of QI efforts. Take the example of the renowned Keystone Intensive Care Unit (ICU) project, which received international attention when it reported a dramatic fall in rates of central venous catheter bloodstream infections in over 100 ICUs in Michigan.[25] Its success has been mistakenly and repeatedly attributed solely to the introduction of a 'simple checklist' rather than a highly complex social intervention.[26] The Michigan programme is likely to have achieved many of its effects through its success in creating a networked community structure that promoted social norms and shared learning. It could therefore be hypothesised that the more attempts to replicate its success rely on single aspects of the programme (such as a checklist), and the more these efforts acquire features of a hierarchy (command-based rather than cooperation-based), the less likely they are to reproduce the original characteristics that contributed to its effectiveness. Cargo cult implementation – it looks like the programme, but it is really not the programme – may explain many of the difficulties that have been experienced in rolling the project out on a wider basis.[27]

This is important, because when an intervention does not work in a new context despite having worked in a demonstration project, there is a danger of mistaking problems of **programme implementation** for problems of **context**. It might be assumed, for example, that the new context was ill-suited to the intervention or incapable of supporting it, and thus abandoned. This means that there is a risk of throwing out interventions that are in fact likely to improve quality of care, based on false assumptions about the interventions.

The problem of describing QI interventions

A sound, full, explicit and theoretically grounded account of QI interventions is clearly indispensable. However, achieving a good understanding of what an intervention **is** and how it works is far from straightforward, and this is where the comparison with the drug development model begins to show strain. Drugs can be specified in precise pharmacological terms, and their causal mechanisms, even if complex, can often (though not always) be reasonably neatly described. Describing a QI project and explicating its mechanisms may be far more challenging.

At a minimum, an explicit description of the components and activities of the QI programme should be produced. Such descriptions are often absent in published reports of QI. But the challenge is more

fundamental than providing a shopping list of what was planned to be done when. Those running programmes may not agree what the programme comprises, may have only a poor grasp of what is going on, or may be obliged to articulate claims for the programme that render it acceptable to various stakeholders but have little to do with how the programme really works or is intended to work. Because of the nature of social interventions, what people implementing a programme say they will do may be quite different from what they do, perhaps because it is very difficult to do what they originally proposed, or because their ideas and actions evolve over time as they learn from their experiences in trying to implement the programme.[28] Decades of evaluation science, particularly in the theory-based evaluation (TBE) tradition, have also taught us that many programmes involving a social or behavioural component have an irreducible tendency to adapt and mutate as the programme proceeds.

For these reasons, contrary to Pawson and Tilley, programmes are not 'theories incarnate', where incarnate means (by dictionary definition) 'turn into concrete form'. Pawson and Tilley argue that:

> *'Programmes are [...] shaped by a vision of change and they succeed or fail according to the veracity of that vision. Evaluation, by these lights, has the task of testing out the underlying programme theories. When one evaluates realistically one always returns to the core theories about how a programme is supposed to work and then interrogates it – is that basic plan sound, plausible, durable, practical and, above all, valid.'*[16]

Yet even identifying the 'basic plan' is often not easy, and **whose** version of the programme theories and visions is to be tested out is far from straightforward.

I propose that QI programmes, initiatives and activities are **what actually happens**, not a manifestation of a theory. In the same way, the behaviour of a drug in the body is what actually happens to the drug, not the concrete realisation of a theory about that drug. What actually happens in social and behavioural interventions – the activities actually undertaken, the emphasis placed on different components, the properties holding the programme together – may bear only a limited resemblance to a formal logic model or protocol specified at the outset of a programme. The importance of focusing on **what actually happens** (in so far as it is possible to access and describe this) is vital because without this there is no possibility of understanding the programme components, explaining how the programme worked, or learning about the contextual influences that buffer or modify programme effects.

The role of practical wisdom in getting QI to work

One of the reasons for focusing on what actually happens in programmes, like QI, that have social and behaviour dimensions, is the role of **practical wisdom**. I want to suggest that practical wisdom is important in QI first in getting programmes to work (and therefore is implicated in what actually happens), and second in studying programmes (recognising and understanding what actually happens). When QI initiatives work, they often do so because practical wisdom is deployed both in the design and running of the programme. If, for example, a programme demonstrates dynamic properties – such as adjusting programme components in response to feedback from participants, and creating bespoke versions of the programme to suit local contexts – it may be precisely practical wisdom that gets the programme to work. Practical wisdom is likely, for example, to have been a critical element of the success of the Michigan programme mentioned earlier. Properties of responsiveness, improvisation, dynamic adaptation and focusing on enabling of participants were just as much part of the programme as activities specified in the project protocol. A focus solely on the formal components leads straight to the cargo cult problem.

Practical wisdom is an idea dating back to Aristotle, and has been more or less continuously rediscovered and renamed ever since, often as forms of practical rationality, practical reasoning, and tacit knowledge. Baumard[29] helpfully summarises the distinctions made by Greek philosophers between four different types of knowledge as follows:

- **Episteme**: abstract generalisation, the kind of universal knowledge that is shared and circulated, taught and preserved. It can be seen as knowledge **about** things.
- **Techne**: the capability and capacity to accomplish tasks.
- **Phronesis**: practical and social wisdom, which is the result of experience and social practice. It is singular and idiosyncratic, acquired by trial and error, and cannot be shared easily.

– **Metis**: conjectural knowledge, which is unpredictable and intuitive. It is like a kind of cunning, that uses ruses, shortcuts, and other tactics to get results, and is embodied into purpose. Like phronesis, it is complex, tacit and difficult to communicate.

The idea of 'metis' is put to use in James Scott's book *Seeing like a state*.[30] The book has nothing to do with healthcare. It concerns top-down interventions by states into complex social systems, where such interventions are assumed to be guided by scientific rationality. Scott argues that where interventions involve 'thin simplifications' of the reality of the systems in which they are being introduced, they may erupt into disaster or end in failure. He discusses examples such as 'scientific forestry' in the 19th century (which created monocrop forests vulnerable to pests and storm-felling), and the 'villagisation' of tribal peoples in Tanzania (which was catastrophic for range conservation and pastoral livelihoods, as well as encouraging cholera and livestock epidemics). Scott sees part of the problem of such interventions as lying in hubris about the superiority of scientific knowledge and a corresponding under-valuing of insider, local, experience-based, contextual knowledge (metis). Those in possession of metis - which Scott defines as involving 'a wide array of practical skills and acquired intelligence in responding to a constantly changing natural and human environment',[30] have the ability to adjust and improvise in response to the complexities of dynamic situations. Metis is, Scott suggests, the form of ad hoc reasoning best suited to complex social tasks where the uncertainties are so daunting that intuition and 'feeling the way' is most likely to succeed.

Scott's argument that some practical choices cannot be adequately and completely captured in a system of universal rules has some evident parallels with quality improvement. Metis is 'plastic, local and divergent', and one of its key strengths is that it allows contextually appropriate adaptations to be made by mobilising local knowledge. Metis can be indispensable both to those designing and leading QI programmes, and to those implementing QI locally. It is thus crucially implicated in enabling context to be taken into account.

QI programme leaders can certainly draw on the epidemiology of known risk factors for QI programmes when they are designing and running their interventions. But they need metis to be able to recognise what is important and relevant about context for their programme, at the multiple different levels at which context is likely to be important. For example, neo-institutional sociology is now teaching us that institutional context is likely to be critical. Institutions include not only formal organisations and structures (for example, the law courts, insurance and payment systems, hospitals) but also non-codified, informal conventions and collective scripts that regulate human behaviour.[31] Institutional structures mediate the extent to which mechanisms of change can be activated, and thus help to explain variability in outcomes. A programme that was successful in a country with a third-party payer system may well encounter stony soil in a system based on national insurance, or in a country where a recent and similar programme ended in bitter failure and recrimination. Practical wisdom is required to apprehend the significance of institutional context.

Practical wisdom is also needed to identify the 'initial conditions' for a QI effort. Many of these are likely to be historically contingent and have a profound impact on what happens to the intervention. Ansell and Gash[32] for example, show that conditions present at the outset can have a critical impact on the ability of collaborative efforts to succeed. Three, they suggest, may be especially important: imbalances in the resources or power of different stakeholders; the incentives that the stakeholders have to collaborate; and the history of conflict or cooperation among stakeholders. Initial starting conditions can help to explain some of the variability seen across organisations participating in QI interventions both in the extent to which they persist with their efforts and in the outcomes they achieve. This can result in massive variability in success. The sociologist RK Merton famously drew attention to the 'Matthew effect':[33] initial advantage begets further advantage, and initial disadvantage begets further disadvantage. For example, prestigious scientists and institutions tend to attract more attention and resource, thus accumulating further prestige. The overall effect is to amplify inequalities. Merton comments that

> *'initial comparative advantages of trained capacity, structural location and available resources make for successive increments of advantage such that the gaps between the haves and the have-nots in science (and other domains of social life) widen until dampened by countervailing processes'.*

The job of the QI leader is to recognise comparative disadvantages and provide the countervailing processes needed to correct for these.

I must emphasise that QI programme leaders do need to be clear about the principles or activities of programmes that should be invariant – the scientific principles of infection control, in the case of the Michigan programme, for instance. Some elements of QI need to be highly standardised, and there is no getting away from that. But at the same time, those leading QI can engage the wisdom and resources of the community of participants (metis) to make local customisations that increase the chance that the programme will work **here**, even if being implemented in a rather different way from how it is being done **there**. In his book, Scott gives the example of the captain of a large passenger ship, who typically turns over control of the vessel to a local pilot to bring it into the harbour because the local pilot has the local contextual skills and knowledge to get the berthing of the ship right in that particular location. A simple example of this in the Michigan project was that each participating unit came up with its own version of the checklist for good practice in central venous catheter insertion: every unit had a checklist, and every checklist contained the same minimum checks, but each one was different because each drew on the practical skills and knowledge of local participants about what was likely to work where they were.

Metis is generally indispensable to the dealing with challenges known locally – who is the doctor likely to create a ruckus about being asked to do this; what are the levers for getting management to authorise the budget; and which individuals and committees will need to be consulted to ensure harmony, for example. Recognising the place of metis in running programmes helps to avert thin, formulaic simplifications of interventions that are likely to lead to disappointment. It is consistent with a recent turn within management theory towards regarding some aspects of quality improvement as more like an art, and requiring qualities of flexibility, dynamism, and creativity that a purely standardised approach cannot hope to achieve.[34]

This has a number of implications.

- First, a proper understanding of **what a QI intervention is** needs to be at the right level of specification. It cannot be derived from an inspection of formal specifications, any more than what an organisation is really like can be derived from an organisation and management chart. It needs to include a role for practical wisdom.
- Second, QI efforts need to accept that, in contrast with drugs, its interventions are never likely to be completely standardisable and fully specified, and that this is indeed desirable. Many aspects of a programme will be immutable, but some aspects will forever escape reduction to a set of executable instructions. QI efforts will always involve trade-offs between explicitness and flexibility if they are to work.
- Third, understanding **what a programme is** (rather than what its designers or other stakeholders think it is) requires studying it in action. This is where the second use for practical wisdom comes in.

The role of practical wisdom in studying and understanding QI activities

I would suggest that when an intervention that worked before does not work when moved to a new context, then:

1. it did not work in the first place (the observed improvement in that first place was really due to something else), or
2. the intervention in the new place is not the same as the intervention in the first place, even if it bears a seeming resemblance (there is heterogeneity in implementing the intervention), or
3. the new place is so different to the first place that the intervention cannot work, or can only work much less effectively (there is heterogeneity in the context), or
4. some combination of 2 and 3 has occurred.

In order to understand which of these applies, there is an equally important role for practical wisdom in **studying** QI initiatives and gaining real insight into how they work. Without well-designed, well-informed social inquiry, it is impossible to understand **what actually happens** in QI, it is impossible to identify the mechanisms that link outcomes to inputs, and it is impossible to account for context. This form of social inquiry, I suggest, requires deployment of a range of methods not often found together in the study of QI at present, and use of practical wisdom in interpreting and synthesising the findings and in feeding them back to the people charged with designing and implementing programmes.

Finding out what actually happens in a QI programme is no mean feat. It is most likely to involve ethnographic methods, including observations of programme team meetings, programme events, and programme implementation at the sharp end; analysis of documents; and interviews with those involved or affected by a QI intervention. Conducting such work across a range of contexts can enable rich insights into the extent to which what is happening conforms to the designers' expectations, and what explains any deviations. It offers the ability to identify the contextual influences on the capacity and willingness of organisations, teams or individuals to implement the initiative – the awkward clerk, the absence of a functioning IT system, the depressed consultant, the history of many previous failed attempts to solve the same problem, the 'normalisation of deviance'[35] that means that people in a specific context are falsely reassured that the problem facing them is not really a problem at all. And, used wisely, such work can be used formatively to feed back directly into the programme, and enhance the wisdom of the programme leaders while the intervention is running. There are still far too few examples of this kind of study in QI, and some of the major methodological and ethical issues have still to be resolved.

Constructing explanations that get inside the black box of causation and that account for context is the next critical, and linked, task for social science inquiry in QI. There can be little doubt that epidemiological studies of the contextual modifiers of QI interventions are badly needed, not least so that those implementing QI interventions have better risk-assessment tools to use. Some of these models may, as I suggested earlier, benefit from the increasing sophistication of statistical techniques now appearing in the clinical sciences. But a science of causation and context cannot be built on such models alone: correlation is not causation, and even though correlational work is indispensable to theory building, a full understanding of what gets a programme to work will elude measurement.

Before discussing this further, it is perhaps worth acknowledging the explosion of the literature devoted to mechanism-based approaches to theory building in the social sciences. Realist evaluation is just one example among very many; Hedstrom and Swedberg's book *Social Mechanisms: an analytic approach to social theory,*[36] published around the same time as *Realistic Evaluation,* for example, makes the same argument as Pawson and Tilley in suggesting that any understanding of mechanisms cannot be derived from correlational analysis alone:

> *'Assume that we have observed a systematic relationship between two entities, say I and O. In order to explain the relationship between them we search for a mechanism, M, which is such that on the occurrence of the cause or input, I, it generates the effect or outcome, O. The search for mechanism means that we are not satisfied with merely establishing systematic co-variation between variables or events: a satisfactory explanation requires that we also be able to specify the social "cogs and wheels" that have brought the relationship into existence.'*[36]

The mechanismic literature demonstrates surprisingly little consensus on what might constitute a mechanism, however.[37] Most social phenomena, as Diego Gambetta points out,[38] require more than one mechanism to explain, but mechanisms do not simply pile up on top of one another. Rather, mechanisms interact with each other, forming what Gambetta terms 'concatenations of mechanisms'.

I am inclined towards the view that discussions of what constitutes a mechanism rapidly become unproductive (and tedious), and that it is often impossible, close up, to distinguish mechanism from context. I prefer to revert to the idea that what social science in QI is about is building middle-range theories. Robert Merton defined mid-range theories at some length,[39] seeing them as lying somewhere between the minor hypotheses used in day-to-day research and attempts to build more all-encompassing 'big' theories of social life. Such theories include a focus on mechanisms, but typically provide a broader narrative.

How we should build mid-range theories in QI seems to me one of the most important challenges. Practical wisdom is needed to interpret the results of ethnographic work and quantitative evaluative research, but it may be that new approaches need to be added to the armoury to ensure the deepest understanding. Some of the most exciting and innovative methodological work is now taking place in the area of case studies. This is beginning to show how the attribution of causality in case studies can be supported by iterative pattern-matching processes that develop explanations, deduce implications of those explanations, and seek

additional information to check these explanations out.[40,41] Charles Ragin's work on fuzzy-set qualitative comparative analysis (fsQCA)[42] is also offering methods that can be used to summarise and order findings from case studies to provide a systematic means of assessing whether causes can reasonably be attributed to effects, and that avoid the pitfalls associated with assuming unit heterogeneity. Much of this work is focused on the identification of **necessary** and **sufficient** conditions for change, and seems to offer a rich source of thinking about context.

4. The principal research questions relating to context

1. What are the best methods for investigating the influence of context on QI activities?
2. Can elements of the social and clinical sciences be blended to produce a framework for the study and implementation of QI?
3. What is the role of pilot studies in clarifying the theories underlying QI efforts and the likely contextual modifiers?
4. How can the toxic effects of QI efforts across different contexts best be assessed?
5. In order to avoid cargo cult QI, can we produce better accounts of what actually happens in QI efforts, and what is the method by which such accounts can best be obtained?
6. Can good epidemiological models of contextual modifiers in QI be built, and can they be used to conduct risk assessments in local settings?
7. Can social science studies running alongside QI efforts provide formative feedback that enhances the ability to adjust for context?
8. How can the role of practical wisdom in running QI programmes be accounted for, and how can current editorial policies in major peer-reviewed journals accommodate it?
9. What is the role of new case study methods in understanding context in QI?
10. What is the best way of synthesising scientific evidence of different types across contexts to produce good programme theories for QI?

References

1 Blumenthal D, Kilo CM. A Report Card on Continuous Quality Improvement. *Milbank Q* 1998;76(4):625-48.

2 Schouten LM, Hulscher ME, van Everdingen JJ, Huijsman R, Grol RP. Evidence for the impact of quality improvement collaboratives: systematic review. *BMJ* 2008;Jun 28;336(7659):1491-4.

3 Øvretveit J, Bate P, Cleary P, Cretin S, Gustafson D, McInnes HM, et al. Quality improvement collaboratives: Lessons from research. *Quality and Safety in Health Care* 2002;11(4):345-51.

4 Grol R, Wensing M. What drives change? Barriers to and incentives for achieving evidence-based practice. *MJA* 2004;180(6 Suppl):S57-60.

5 Black N, Thompson E. Obstacles to medical audit: British doctors speak out. *Social Science and Medicine* 1993;36(7):849-56.

6 Lomas J. Using research to inform healthcare managers' and policy makers' questions: From summative to interpretive synthesis. *Healthcare Policy* 2005;1(1):55-71.

7 Scott I. What are the most effective strategies for improving quality and safety of health care? *Intern Med J* 2009;Jun;39(6):389-400.

8 Goodin RE, Tilly C. *The Oxford handbook of contextual political analysis.* Oxford University Press; 2006.

9 Falleti TG, Lynch JF. Context and causal mechanisms in political analysis. *Comparative Political Studies* 2009;42(9):1143-66.

10 Davidoff F. Heterogeneity is not always noise. *JAMA* 2009;302:2580-6.

11 Shekelle PG, Pronovost PJ, Wachter RM, Taylor SL, Dy S, Foy R, et al. Assessing the Evidence for Context-Sensitive Effectiveness and Safety of Patient Safety Practices: Developing Criteria. (Prepared under Contract No. HHSA-290-2009-10001C). AHRQ Publication No. 11-0006-EF. Rockville, MD: Agency for Healthcare Research and Quality; December 2010; 2010.

12 Pawson R, Tilley N. *Realistic Evaluation.* London: Sage Publications Ltd; 1997.

13 Greenhalgh T, Humphrey C, Hughes J, Macfarlane F, Butler C, Pawson R. How do you modernize a health service? A realist evaluation of whole-scale transformation in London. *Milbank Q* 2009;Jun;87(2):391-416.

14 Newlands ES, Stevens MFG, Wedge SR, Wheelhouse RT, Brock C. Temozolomide: a review of its discovery, chemical properties, pre-clinical development and clinical trials. *Cancer Treat Rev* 1997;1;23(1):35-61.

15 Dixon-Woods M, Cavers D, Agarwal S, Annandale E, Arthur A, Harvey J, et al. Conducting a critical interpretive synthesis of the literature on access to healthcare by vulnerable groups. *BMC Medical Research Methodology* 2006;6(35).

16 Pawson R, Tilley N. Realist Evaluation. In: Otto H, Polutta A, Ziegler H (eds). Evidence-based practice: modernising the knowledge base of social work? Farmington Hills, MI: Barbara Budrich; 2009.

17 Betensky RA, Louis DN, Cairncross JG. Influence of Unrecognized Molecular Heterogeneity on Randomized Clinical Trials. *Journal of Clinical Oncology* 2002;May 15;20(10):2495-9.

18 Hegi ME, Diserens A, Gorlia T, Hamou M, de Tribolet N, Weller M, et al. MGMT Gene Silencing and Benefit from Temozolomide in Glioblastoma. *N Engl J Med* 2005;03/10;352(10):997-1003.

19 Pawson R. *Evidence-based policy: a realist perspective.* London: Sage; 2006.

20 Healy KJ. *Last best gifts: altruism and the market for human blood and organs.* Chicago; London: University of Chicago Press; 2006.

21 Shortell SM, Bennett CL, Byck GR. Assessing the Impact of Continuous Quality Improvement on Clinical Practice: What It Will Take to Accelerate Progress. *Milbank Q* 1998;76(4):593-624.

22 Burton PR, Tobin MD, Hopper JL. Key concepts in genetic epidemiology. *Lancet* 2005;Sep 10-16;366(9489):941-51.

23 Shojania KG, Grimshaw JM. Evidence-based quality improvement: The state of the science. *Health Aff* 2005;24(1):138-50.

24 Feynman R. Cargo cult science: some remarks on science, pseudoscience, and learning how to not fool yourself. In RP Feynman and J Robbins. *The Pleasure of Finding Things Out.* Cambridge, Mass.: Perseus Books; 1999; 205-16.

25 Pronovost P, Needham D, Berenholtz S, Sinopoli D, Chu H, Cosgrove S, et al. An intervention to decrease catheter-related bloodstream infections in the ICU. *N Engl J Med* 2006;355(26):2725-32.

26 Bosk CL, Dixon-Woods M, Goeschel CA, Pronovost PJ. The art of medicine. Reality check for checklists. *The Lancet* 2009;374:444-5.

27 Pronovost PJ. Learning accountability for patient outcomes. *JAMA* 2010;Jul 14;304(2):204-5.

28 Weiss C. *Evaluation: Methods for studying programs and policies.* 2nd ed. Upper Saddle River, NJ: Prentice Hall; 1998.

29 Baumard P. *Tacit knowledge in organizations.* London: SAGE; 1999.

30 Scott JC. *Seeing like a state: How certain schemes to improve the human condition have failed.* Yale Univ Pr; 1998.

31 Mahoney J, Thelen KA. *Explaining institutional change: ambiguity, agency, and power.* Cambridge University Press; 2009.

32 Ansell C, Gash A. Collaborative Governance in Theory and Practice. *Journal of Public Administration Research and Theory* 2008;October 01;18(4):543-71.

33 Merton RK. The Matthew effect in science, II. Cumulative advantage and the symbolism of intellectual property. *ISIS* 1988;79:606-623.

34 Hall JM, Johnson ME. When should a process be art, not science? *Harvard Business Review* 2009;March:58-65.

35 Vaughan D. *The Challenger launch decision: risky technology, culture, and deviance at NASA*. Chicago; London: University of Chicago Press; 1996.

36 Hedstrom P, Swedberg R. *Social mechanisms: an analytical approach to social theory.* Cambridge: Cambridge University Press; 1998.

37 Mahoney J. Beyond correlational analysis: recent innovations in theory and method. *Sociological Forum* 2011;16:575-592.

38 Gambetta D. Concatenations of mechanisms. In: Hedstrom P, Swedberg R, (eds). *Social mechanisms: an analytical approach to social theory*. Cambridge: Cambridge University Press; 1998; 340.

39 Merton RK. *On social structure and science.* Chicago; London: University of Chicago Press; 1996.

40 Mark MM, Henry GT, Julnes G. Toward an integrative framework for evaluation practice. *American Journal of Evaluation* 1999;20(2):177.

41 Mahoney J. After KKV: The New Methodology of Qualitative Research. *World Polit* 2010;62(01):120-47.

42 Ragin CC. *Fuzzy-set social science.* Chicago: University of Chicago Press; 2000.