



**Paul Burton**

**Abstract** Evaluation research is an important part of the broader field of urban analysis. It draws on many of the same techniques and approaches used by other types of research about urban areas and processes, including social, economic and environmental science methods, but uses them to make judgements or evaluations. These evaluative judgements are not based on the preferences or priorities of the evaluators (although they invariably have them), but on criteria established by others. These criteria might be the aims and objectives of an urban policy intervention or the assumptions underpinning a land-use planning scheme. This chapter traces the development of different conceptions and approaches to evaluation, from experimental and constructivist through to realist approaches. It looks also at the particular challenges faced by evaluation researchers looking to judge the success of complex, area-based initiatives which are still a mainstay of much contemporary urban policy. The chapter looks also at the politics and ethics of evaluation research, including the challenges of ‘speaking truth to power’, and recommends the continuous development of the craft skills of evaluation research alongside an ongoing commitment to the fundamental principles of good research design.

## 1 What is Evaluation Research?

As earlier chapters have discussed, urban analysis covers a multitude of perspectives and practices with a common focus on phenomena and processes that have an urban dimension. Of course, this does not necessarily help us in defining what we mean by ‘the urban’ with any precision as many processes such as job creation and loss, housing and community development or travel behaviour have fundamental characteristics that may or may not vary according to their urban setting. Likewise, many phenomena that we might wish to study in an urban setting, such as unemployment, social isolation or anti-social behaviour, can be found also in suburban or rural settings. But, for the time being, let us put to one side the challenge of differentiating

---

P. Burton (✉)  
Griffith University, Gold Coast, QLD, Australia  
e-mail: [p.burton@griffith.edu.au](mailto:p.burton@griffith.edu.au)

© Springer Nature Singapore Pte Ltd. 2021  
S. Baum (ed.), *Methods in Urban Analysis*, Cities Research Series,  
[https://doi.org/10.1007/978-981-16-1677-8\\_10](https://doi.org/10.1007/978-981-16-1677-8_10)

urban phenomena from those generated by or experienced in other types of area and focus instead on evaluation research as a vital element of urban analysis, for while the phenomena in question might vary, the means of analysing them and evaluating policy responses to them is not significantly different.

However, definitional challenges exist also when we consider the field of evaluation and evaluation research, not least in exploring the specificity of the main techniques of evaluation research *vis a vis* social and economic research more generally. What can be said about any form of evaluation, which inevitably requires research in one form or another, is that it is judgemental. This is not to say that moral judgments are made—although they might be—simply that evaluation involves assessing something against a set of existing criteria. As the renowned evaluation theorist, Michael Quinn Patton [1, p. 1] observed in a recent critical summary of the field,

It's all about criteria. Criteria are the basis for evaluative judgment. Determining that something is good or bad, successful or unsuccessful, works or doesn't work, and so forth and so on requires criteria

In other words, while urban analysis is a broad field it includes the similarly broad field of evaluation research, which is an unavoidably judgemental exercise in which degrees of success or failure are judged against criteria. In the urban realm, those criteria could be a plan or strategy, they could be a new planning process, or they could be more tangible outcomes seen on the ground—a new suburb or a park, for example. Evaluation in this sense relates typically, but not exclusively, to interventions that are expected to have certain desirable outcomes, but these interventions can originate from public or statutory bodies such as local or state governments or from private developers be they large corporations or individuals.

Some have argued the case for 'non-judgemental evaluation' [2] but this has always seemed (to me at least) to be a *non sequitur*, a proposition that does not follow logically from its premises. An advocate for non-judgemental evaluation might claim to be providing only a description of the outcomes and impacts from an intervention, while leaving the conclusive judgement to others. But this suggests the initial descriptive work is only part of the evaluation process, an important part to be sure, which is not complete until someone makes a judgement call. It is worth noting that this position is similar to the philosophical distinction made between facts and values and to the practical distinction that underpins the notional relationship between civil servants or state bureaucrats and their political masters. In this case, politically neutral civil servants provide advice to politicians who make decisions based on their clearly articulated political values.

There is, however, a well-established viewpoint within the literature on the theory and philosophy of evaluation [3] that says we should not begin with a set of clearly stated objectives or goals and examine the extent to which they are realised in practice. Rather, we should look at what happens in practice and then try to deduce a set of objectives that would make sense of that practice. This is a logical approach to take, especially if you prefer an inductive rather than deductive approach to scientific

exploration. However, it struggles with one of the same practical challenges of inductive methods, namely that it can be extremely difficult (some would say impossible) to clear one's mind of preconceptions before collecting and analysing empirical data and even to know where to focus one's analytical gaze at the start of this process.

In this chapter, I assume, therefore, that the most productive way to plan and carry out a piece of evaluation research is to look at the stated objectives, aims or goals of an intervention and then to think about how you might go about assessing the extent to which they have been realised. In the course of this and perhaps in conclusion, one might argue that the stated objectives were mainly symbolic [4] and there was never any genuine intention of carrying them out or manifestly insufficient resources were devoted to their implementation, in which case a more covert set of objectives might be described. But the essence of evaluation research involves judging something by the standards it sets itself.

To summarise, the broad field of urban analysis includes evaluation research, or research carried out as part of an evaluation. Much research in the urban field involves the evaluation of urban policy measures, such as area-based initiatives which we will look at in more detail later. But it includes also research carried out as part of the development of urban policy and in the implementation of planning policy, so that planners are able to judge whether or not a development proposal meets the requirements and intent of a planning scheme and can be recommended for approval.

This chapter explores some of the key methodological, conceptual and practical challenges of doing evaluation research in urban settings and about urban processes and phenomena. It does not provide detailed instructions for how to go about designing and conducting this type of research as there are many excellent books on these practicalities, listed at the end.

Finally, it is important to acknowledge the significant differences that can exist between urban analysis or research produced by academics for other academics and urban research produced with the primary intention of being useful to policy-makers and practitioners. Of course, academic researchers can choose whether to direct their work mainly at fellow academics or at practitioners and some choose to serve both audiences, and researchers beyond the academy can likewise choose to present their urban analytical work and/or evaluation research, undertaken initially for practitioners and policy-makers, to academic audiences. There is a growing body of work that examines the relationship between research for these different audiences, exploring the motivations of researchers and their ability to present critical conclusions. While some of this new work offers respectful and thoughtful insights on the challenges of moving between these worlds [5–7], there is another body of work that claims to be critical but is characterised by a tendency to political naiveté and academic piety built on unsophisticated conceptions of the relationship between research, policy and politics [8].

## 2 Methodological Challenges

If much of the substance of urban analysis and evaluation research involves questions of what works, to what extent, and at what cost, then issues of causation are central. Sometimes in evaluation research, we talk of theories of change or program logic when referring to underlying causal mechanisms, but sometimes we find it difficult to identify them at all or we struggle to grasp their complexity and fall back on notions of 'black boxes'. These metaphors allow us to deal with inputs, outputs and outcomes, without understanding the processes by which inputs are turned into outputs that have outcomes. For example, I have a number of black boxes that take inputs of electricity and data stored on a compact disk and turn it into music that I can hear and enjoy. While I know that without those inputs, I would not be able to listen to music from that black box (which might of course be white or silver), I have little or no idea of what goes on inside it: how data are extracted from the disk, turned into music and broadcast to me. I might be able to assess the quality of these inputs (the reliability of the electricity supply, or the significance of scratches on the disk) and know that they are likely to affect the quality of the output and hence my enjoyment of it, but, again, I know little about why a scratch cannot be dealt with by the CD player.

Some evaluation research deals with complexity and a lack of understanding of causal mechanisms by using a similar black box metaphor and focussing instead on more tangible and measurable inputs and outputs. I was involved in various evaluations of UK urban policy initiatives in the 1980s and 1990s, commissioned by the government departments responsible for them, in which the emphasis was almost entirely on the measurement of inputs and outputs, with only occasional reference to outcomes and rarely any consideration of causal mechanisms. As noted in the introduction, many of these took the form of area-based initiatives in which selected areas would receive additional funds and sometimes enhanced powers to alleviate local manifestations of poverty, deindustrialisation, environmental degradation and poor local public services. As part of the evaluation of these programs, we would scrutinise budgets, check that money allocated to local groups to deliver services such as childcare, environmental improvements or job training was properly spent and then monitor certain outputs. These might include the number of new childcare places available, the number of trees planted, or training sessions delivered. Sometimes, our measurements would be rather more sophisticated and we would track whether the childcare places enabled parents to look for work or have the time to attend job training sessions, or whether the saplings planted ever managed to grow into established trees or whether the job training sessions helped people get jobs. Because each of these measures in themselves provided only a very partial picture of the anticipated transformation of an area, a summative assessment was also important but rarely called for or delivered. As we will explore in more detail below, there are considerable methodological problems in establishing robust causal connections between intervention measures and outcomes, but that is no good reason for not attempting to do so. This more rounded evaluation would also scrutinise the

assumed causal connections and mechanisms that lead policy-makers to assume (in this case) that a lack of childcare places inhibit the labour force participation of parents of small children, or that planting trees helps restore a sense of pride and wellbeing into residents or that job training enhances employability. Without this focus on causal mechanisms, we are not well placed to claim that we have good theories of causation, such that we can say with some confidence that a combination of policy intervention X, Y and Z produced this particular outcome for this place. As we will consider below, a variety of factors often combine to make this type of summative evaluation very difficult to carry out.

But first, an important digression into the philosophy of knowledge and science and a relatively recent development that has had a profound impact on the design of robust evaluations. While philosophical realism can be traced back to Plato, it was not until the 1960s that works now considered seminal in the field began to appear—Rom Harré's *Introduction to the Logic of the Sciences* [9] and Roy Bhaskar's *A Realist Theory of Science* [10] being good examples. Critical realism is in essence the belief that causal mechanisms exist and account for why certain things happen. The 'things that happen' are real and not simply figments of our imagination and can be captured or described empirically. This position developed and extended its reach into the field of evaluation for a number of reasons, not least because of a growing disquiet with on the one hand the excesses of constructivism, especially when combined with impenetrable postmodern jargon, and on the other a rejection of the simplistic and a theoretical assumptions of empiricism.

In 1997, Pawson and Tilley's *Realistic Evaluation* [11] provided the first comprehensive and compelling summary of how to apply the principles of critical realism in the evaluation of public policy measures. The opening chapter of their book presents (in 28.5 pages) a brief history of evaluation over the preceding 30 years, including what had become known as 'the paradigm wars' between advocates of experimental approaches, pragmatists, constructivists and latterly, realists.

The experimental approach is rooted in a common-sense notion of causation that relies on comparing similar or identical groups (perhaps of people or places), one of which receives an intervention (an expedited planning regime for example) while the other does not. By measuring the volume and throughput of development before and after a given period of intervention (perhaps one or two years in this case), we then assume that any difference is attributable to or caused by the intervention. While this approach can be made to work reasonably well in situations like a laboratory where all the possible variables or characteristics of the two groups can be controlled, in the field or 'real world' this is not only more difficult, it is perhaps impossible. Some have argued that policy reforms or interventions that are selective rather than universal allow for comparisons that are close enough to the assumptions of classical experimental design to be worthwhile [12–14] and more recently a number of governments, including in Australia and the UK have established new policy research teams that use experimental approaches and behavioural economics to provide more robust empirical evidence to policy-makers.

In the face of a growing volume of evaluation research that struggled to find any definitive answers to the question of 'what works?', a new strand of pragmatism

emerged among evaluators. This made the assumption that the practice of policymaking and policy implementation is complex and messy rather than rational and coherent and, therefore, evaluation research needed to adopt a more nuanced approach, which recognised that policies and interventions will change over time and that evaluation research will be only one factor in this turbulent environment. It also recognised that the value placed on research and its conclusions is a messy process that was, to some extent, a product of political values and preferences as much as any claim to objective truth. In other words, a piece of evaluation research might be more influential in policy circles if its conclusions aligned with the priorities of leading politicians and bureaucrats regardless of its methodological rigour or epistemological purity.

This movement away from absolute notions of truth and objectivity reflected debates in the much wider field of epistemology and the philosophy of science and social science. But more importantly, it reflected a recognition and acknowledgement that evaluation research is unavoidably part of a political process and that the power relationships between those who commission and those who conduct an evaluation are significant. Critical commentators argued that the principle of ‘whoever pays the piper, calls the tune’ must be recognised as it calls into question some claims to the maintenance of professional standards and objectivity. We return to this below when considering the challenge of ‘speaking truth to power’. During the 1980s and 1990s, the constructivist conception of evaluation held sway [15] wherein evaluators negotiate or construct what might now be termed an assemblage of perspectives and views held by the full range of stake-holders in any intervention or policy regime. Critics of this position point to the inequalities of power held by this range of stakeholders and its consequences for trying to synthesise and reconcile these views and opinions and even Guba and Lincoln [15, p. 45] recognised some of the conceptual challenges of this approach when they said,

Evaluation data derived from constructivist inquiry have neither special status nor legitimation; they represent simply another construction to be taken into account in the move towards consensus.

A variant of this recognition underpinned the development a more pluralist conception of evaluation popularised by Rossi, Freeman [16]. This set out a more catholic stance in which it was assumed to be possible, indeed necessary, to accept and combine a variety of approaches into a more comprehensive frame. As Pawson and Tilley [11, p. 24] put it,

One can imagine the attractions of a perspective which combines the rigour of experimentation with the practical nous on policy making of the pragmatists, with the empathy for the views of stakeholders of the constructivists.

By avoiding the thorny question of whether fundamentally different approaches to research and indeed conceptions of knowledge can be combined, the pluralist approach did help evaluators by suggesting that in practice, different aspects of an intervention might be best understood through different research approaches, even if coming to a conclusive synthesis remained a practical and conceptual challenge.

Finally, we see the emergence of theory-driven evaluation in the work of Chen and Rossi [17] and Chen [18], which seeks to explain not just whether an intervention appears to work, but why it does so. An important element of this approach, which contrasts with one of the fundamental principles of the classical experimental approach, is the treatment of variables that might explain differences in outcomes. While experimentalists typically look to isolate the one factor or variable that might explain success (or failure), the more modern theory-driven approach recognises that a number of variables might help explain why an intervention works well in one place at a particular moment in time, but not elsewhere or at a different moment.

This then served as a foundation on which Pawson and Tilley built their conception of realistic evaluation. They start with a notion of theory that contains three elements, linked in ways that help us understand the processes of causation. The first, causal mechanisms (M) are plausible accounts of why something might produce an effect because of the connections between them. The second are contextual factors (C) that influence the extent to which these causal mechanisms come into play and the third are the outcomes (O) produced by this combination of causal mechanisms and contextual factors. They summarise this as:

$$M + C = O$$

Consider this example from urban policy: a local authority constructs a small factory and workshop space in the belief that it will stimulate (cause) local economic development through the establishment or growth of small businesses. However, if the rents charged or lease terms are not sufficiently attractive (contextual factors) then the objective or outcome of local economic development will not be realised. From this, we should not necessarily conclude that the theory of building small factory and workshop units to stimulate local economic development is flawed, but rather that to work effectively it requires some other contextual factors (such as suitable rents and lease terms) to be in place. Hence advance factory building might work very well in one place, but not in another. An evaluation based on one case study would struggle to reach this conclusion, while one that involved a lot of similar interventions in different settings (contexts) would enable this more sophisticated conclusion to be reached.

### 3 Evaluation Research and Urban Analysis

Having explored some of the difficulties encountered in conceptualising evaluations in general, we can now focus on urban policy evaluation, bearing in mind my claim that the distinctiveness and specificity of the ‘urban’ remain unclear and contested. And we need to be wary of suggestions that evaluation research carried out as part of a broader process of urban analysis draws on a separate set of techniques and approaches. But what is the domain of urban analysis?

Urban policy initiatives have for many years, in Australia and elsewhere, involved place-based interventions. This approach has many features that are attractive to policy-makers: it allows resources to be targeted to particular areas where problems are greatest and especially inter-connected, and it enables experimental approaches to be explored and modified before being applied more widely. But there are problems as well, not least that it requires a high degree of institutional coordination and cooperation in practice as well as in rhetorical commitment and it has to cope with the challenges associated with the notion of the ecological fallacy. This recognises that characteristics that apply at the group, or in this case neighbourhood scale, will not apply to all or even the majority of individuals living in that area. In the case of a place-based urban policy measure that helped every firm in a particular neighbourhood on the grounds that overall firms were struggling, some might not be struggling but receive this assistance, nevertheless. This undermines the claim that area-based initiatives are especially effective at targeting scarce resources where they are most needed.

Another more serious set of challenges to researchers responsible for evaluating this commonplace form of urban policy was described many years ago by Robson and Bradford [19] who identified a number of conceptual problems they called 'the six Cs'.

The counterfactual problem is one the most difficult and asks, 'what would have happened without this policy intervention?'. In laboratory settings, the traditional way of addressing this problem is to use an experimental design in which a sample of people who are otherwise similar are randomly allocated between two groups, one of which receives the intervention (a drug perhaps) while the other receives a placebo or dummy drug. Because the recipients are otherwise similar then any differences in experience or outcome are attributed to the effect of the drug. While this research design is seen by many as the gold standard in rigorously evaluating impact or effect, it has also been subject to sustained and considerable criticism. Although theory and practice are constantly evolving, it is still difficult to select 'policy on' and 'policy off' areas in the neighbourhood or suburb scale that can be managed in ways that achieve the requirements of a classic experiment.

Confounding factors include other measures that might produce similar effects to those under investigation in an urban policy evaluation. For example, in a particular location, there might be a policy measure that provides exemptions from certain business taxes (rates, payroll tax, etc.) for local business in order to support and stimulate their establishment or expansion. During the course of this program, interest rates set at the national level might decrease significantly and affect the operational costs of all businesses, including those in the intervention location. If local businesses are seen to be doing well or better than usual in this period, it can be difficult if not impossible to attribute this to the local tax concession or to the national interest rate change. Unlike in a laboratory setting, it is rarely possible to control any possible confounding variables in order to isolate the effect of the intervention in which we are interested.

Contextual factors describe the peculiarities or specificities of particular settings and are important in helping explain why general descriptions or explanations might



vary because of these peculiarities. This distinction has been developed and applied in the most sustained and sophisticated manner by proponents of realistic evaluation such as Pawson and Tilley (1997) and more recently by members of the RAMESES projects ([http://www.ramesesproject.org/Home\\_Page.php](http://www.ramesesproject.org/Home_Page.php)). As we noted above, the corner-stone of realistic evaluation is the formula: Outcome = Mechanism + Context, which represents both a critique and an extension of the basic question that underpinned evaluation research undertaken as part of the evidence-based research movement, namely ‘what works?’. The basic question for realist evaluators became, ‘what works in which circumstances and for whom?’. In other words, locally variable circumstances or contexts will affect local outcomes, even if there is a more general underlying causal mechanism at work. Robson and Bradford [19] illustrate this by noting that in selecting local places or neighbourhoods for an urban policy intervention, there will be significant differences in local context that must be considered when evaluating the impact or success of that intervention. These differences might include the social composition of the area and how it has changed over time, its history of economic development and the tradition of community development and political engagement.

Contiguity challenges relate to what are also known as ‘spillover’ and ‘shadow’ effects. Spillover effects, as their name suggests, occur when the benefits (or indeed the negative effects) of an intervention are not confined to the target area but spill over into surrounding areas. This can present challenges to the evaluation researcher who has to decide how far afield to go in search of measurable impacts, be they positive or negative. Shadow effects operate in the other direction of causality and can be similar to the confounding effects described above, although they typically are more localised. For example, a major employer located just outside a target area might close and have a consequential impact on small firms within the target area that previously supplied goods to the major firm and services to its workers.

These contiguity challenges, along with those presented by variable contextual factors, are associated with all area-based policy initiatives because they rely on drawing boundaries that include and exclude places and people, sometimes on a fairly arbitrary basis.

Combinatorial problems arise when slightly different mixes or combinations of urban policy measures are applied in the areas targeted for intervention. Because these combinations are not typically described or measured with any degree of precision, it is difficult to attribute any variability of overall success to differences in combination or to other factors. For example, while 10 areas might be selected for an urban intervention ‘package’, each of these ‘packages’ might involve a different mix of measures, such as rate relief for local firms, employment subsidies for hiring local workers and environmental improvements. One might offer 50% rate relief for twelve months, while another offers 30% relief for 36 months; some might rely on tree planting to improve the local environment while others deploy a rapid response graffiti cleaning service. In this situation, while all of the targeted areas enjoy the similar status of being an Urban Policy Pilot Area (for example), each is implementing this program in slightly different ways and the evaluation researcher is faced with yet another significant problem of attribution.

Changes over time present the final challenge to evaluators. It is commonplace to argue that place-based urban policy interventions designed to solve complex and multi-faceted problems cannot be expected to solve significant problems in the short term, which might reasonably be taken to mean 2–3 years. Many have argued that these interventions must be left in place for decades if they are to have any chance of achieving profound and transformational improvements [20]. If this long-term commitment is achieved, and it must be said that most programs of this type are either abandoned or radically transformed after a few years, then the evaluator has to have a framework that takes account of changes over time. These changes will be complex and difficult to analyse. For example, some measures might be implemented immediately as they are seen as foundational for the success of other measures and then phased out having done their job. For example, environmental cleanups or crime prevention measures might be seen as a necessary precursor to the introduction of more direct economic stimulation measures, but measures that can be wound up or wound down once they have achieved short term impacts. Or, they might be abandoned because they are judged (by interim evaluation) to have not been successful. Again, there is often significant variation in the pattern and timing of these changes between different target areas and for good reason. What works in one area, might not work so well in another. Capturing this variation and building it into a comprehensive analysis of what works is difficult but essential to any rigorous evaluation.

While Robson and Bradford [19] identified these methodological challenges in designing rigorous evaluation frameworks for complex area-based initiatives some years ago, they are still pertinent today, not least because so many urban policy interventions retain their focus on select and closely defined localities. Interestingly, some of the measures brought into counter the spread and impacts of the COVID-19 pandemic have again drawn attention to the logic of spatial targeting and the boundaries used in particular instances. While popular debate rarely uses the term, it is often about the ecological fallacy whereby the assumed characteristics of a given area or suburb are taken (erroneously) to apply to everyone living in that area. We are increasingly aware that spatial targeting should be used with great caution as part of any urban policy intervention and that new datasets containing much more individually tailored information might be more appropriate in focusing policy measures on those who most need them.

## **4 Evaluation Research as Part of Urban Planning**

In addition to the plethora of research carried out as part of urban policy development and evaluation, we should not forget that statutory land use planning as practised around the world, typically involves a significant research and evaluation component, even if it is not always seen in this light. Planners carry out research on urban conditions as part of the planmaking process and incorporate many assumptions, for example about future populations, that are subject to evaluative scrutiny over

time. More significantly, development proposals and applications are assessed or evaluated against increasingly complex sets of criteria that are enshrined in local planning schemes and strategic plans. While not often recognised as such, this is probably the most significant form of evaluation research that is undertaken on a day-to-day basis by planners. Development proposals are assessed in terms of their environmental and social impacts and recommended for approval or rejection against the criteria contained in the planning scheme. This typically requires the planner to either conduct the assessment themselves or to commission and use a specialised assessment from another expert: in ecology, hydrology, soil mechanics, acoustic engineering and so on.

There is, however, another way in which planners carry out evaluation research. Planning theory sometimes distinguishes between procedural and substantive theory, in other words between theory about the processes adopted by planners, such as comprehensive or inclusive approaches when going about their business and the substantive issues they focus on, such as community development, creating vibrant public realms, ensuring transport systems work well or maintaining housing standards. In this respect, evaluation research in planning tends to focus on the substantive interventions of the planner and there has been a dearth of research that evaluates some of the key procedural claims and assumptions about planning, such as the benefits of greater public participation [21, 22].

## 5 The Politics and Ethics of Evaluation Research

Because evaluation research can never be simply an exercise in the application of technical skills, but inevitably requires judgements to be made and values to be applied, it is a process imbued with politics. This is not to say that evaluation research is necessarily a partisan activity, in which ‘experts’ associated clearly with different political groups or parties compete to have their assessment accepted—although this of course happens. Rather, it suggests that judgements involve choices, and these have consequences that affect different groups or individuals in different ways. Understanding and appreciating that one is working in a political environment should lead the accomplished evaluation researcher to be aware of the choices they make, including those that appear to be simple methodological choices, and their consequences. This is not to say that important principles should be compromised, but that practitioners may well be confronted with challenges to their principles and should be self-conscious in upholding them or otherwise.

In the remainder of this section, we explore some of these challenges through three distinct but related questions:

- Can researchers be objective and neutral in their work?
- How should we learn from mistakes?
- Can and should evaluation researchers speak truth to power?

## 5.1 *Can Researchers be Objective and Neutral in Their Work?*

While absolute objectivity and neutrality—or perhaps impartiality is a better term—might be unattainable, evaluation researchers can choose to make a commitment (at least to themselves) to conduct an evaluation that could in principle conclude with a position they do not like or prefer. One could argue that if an evaluator is unwilling to make this principled commitment, they should not accept the commission of that evaluation. Similarly, if evaluation researchers have reservations about the ways in which those commissioning an evaluation might use their work, they should consider carefully whether or not to accept or bid for the commission. These reservations might include having control over the design of an evaluation, having access to all relevant data, being able to present conclusions honestly reached, having some control over how the results are published and presented, and being able to publish results separately—perhaps in academic journals.

As with any commission, the more clarity that is provided at the outset about these ‘terms of engagement’, the better. But disputes might still arise around one or more of these aspects which will test the integrity and good faith of all parties. Developing a thoughtful and comprehensive contract helps in these processes and can be seen as a sign of evaluation becoming more professional in its approach.

Rossi et al. [23, p. 404] speak of evaluation becoming a more ‘professionalized’ field, even if it has not become a profession in the widely accepted sense of the word. Part of this process involves being able to identify and adhere to certain standards when conducting an evaluation. While there is no universally accepted set of standards that must be adopted by any practising evaluator or evaluation researcher, common standards typically include:

- A commitment to systematic, rigorous and perhaps evidence-based work;
- A promise of a degree of competence and experience in conducting evaluations that are systematic and rigorous;
- An expectation that evaluators are honest in their work, not falsifying data, wilfully misinterpreting it or hiding data that does not support the conclusions reached;
- A promise to engage with program managers, clients and other stakeholders in a respectful manner;
- A commitment to framing evaluation in relation to some notion of the public interest or common good.

Of course, each of these principles can be and are subject to criticism. For example, there is now an abundant literature that takes issue with the often simplistic assumptions of evidence-based policymaking and practice (eg., Shaxson and Boaz [7] and Burton [24]), especially in its earliest manifestations. It can be difficult to disentangle concerns with the conduct of an evaluation from dislike of its findings and conclusions and respect for program workers might be difficult to maintain if an evaluation uncovers widespread fraud or egregious behaviour. As in some of our earlier

discussion, these principles are in many respects not specific to evaluation practitioners, evaluation researchers or urban analysts. They are fairly universal and indeed the United Nations Educational, Scientific and Cultural Organization (UNESCO) has published an updated Recommendation on Science and Scientific Researchers, which sets out a series of similar statements about how research should be undertaken [25]. To drive the implementation of this, a global network—the Responsible Research and Innovation Networked Globally (RRING)—has been formed with the following mission:

We are a coalition that has activism at its core. We seek to make research and innovation systems everywhere more responsible, inclusive, efficient and responsive as an integral part of society and economy. (<https://rring.eu/>)

But as many program evaluators discover through their own work, there can be a significant gap between the statement of program goals and objectives and their achievement in practice and it remains to be seen whether or not statements such as these succeed in raising the standards of actually existing evaluations or remain as mere statements of good intent.

One of the biggest and longest-standing challenges to the integrity of evaluation research lies in the commitment to speak truth to power. While Aaron Wildavsky's [26] classic text, *Speaking Truth to Power: The art and craft of policy analysis* popularised the term at a time when policy studies were emerging as a distinctive discipline, the expression (*parrhesia* in Greek) first emerged in the writings of Euripides and was an essential component of Athenian notions of democracy [27].

## 5.2 *How Should We Learn from Mistakes?*

The Behavioural Economics Team of the Australian Government (BETA) represent another relatively new phenomenon in some governments—they exist to apply behavioural insights into contemporary policy problems, often using the tactic of making small changes or ‘nudges’ to address longstanding problems. While more seasoned policy scientists might see this as a modern version of Lindblom's incrementalism, in this manifestation one of the more interesting elements of the work of the BETA within the Department of Prime Minister and Cabinet is their stated commitment to learning from mistakes and to publishing the results of their work, even if it shows that particular interventions were not as successful as hoped for. While there are no readily available assessments of how this principle is applied in practice, it represents a commendable position for any government entity to take.

In a similar vein, some governments—national as well as state or provincial—have created new positions, such as Government Chief Scientist, Chief Economist or Government Statistician to oversee and ensure the integrity of research carried out by or on behalf of the government. In the UK, the Government Office for Science published recently *Guidance for Government Chief Scientific Advisers and their Officials* to clarify roles and responsibilities.

### ***5.3 Can and Should Evaluation Researchers Speak Truth to Power?***

In many OECD countries, especially those that adhere to the Westminster system of government, elected politicians are supported in their roles by professional public servants and while under threat in recent years, this public or civil service has been relatively permanent. The character of Sir Humphrey Appleby in the BBC TV satirical comedy *Yes, Minister*, is a Permanent Secretary, meaning he remains while the ministers he serves come and go according to electoral cycles and factional politics within the parties of government. The underlying logic of this arrangement is that these public or civil servants are able to offer ‘frank and fearless’ advice to their political masters without having to worry that they might lose their jobs by suggesting to a minister that his or her preferred policy has some logical flaws or serious impracticalities and is unlikely to achieve its objectives.

This approach of being frank and fearless or speaking truth to power remains an important part of the discourse of government, even as evidence mounts that the public service in Australia and elsewhere are being systematically politicised to the detriment of sound policymaking and good outcomes [28].

But of course, the assumptions of professional objectivity and the commitment to speak truth to power have also been subject to sustained criticism and indeed the very success and popularity of *Yes, Minister* lay in its portrayal of Sir Humphrey and his fellow mandarins as powerful people in their own right, with policy agendas of their own, even if these are small-c conservative and not especially partisan. Equally insightful and funny portrayals of the relations between politicians and their public servants in Australia can be seen in the ABC TV programs *The Hollowmen* and *Utopia*, programs that some public servants have said are far too close to their day-to-day worlds to be watchable.

The point of referring to these popular portrayals of the relationships between public servants who are democratically elected and those who are appointed because of their professional skills and scientific or technical knowledge is to provide some context for the environment in which evaluators, urban analysts and urban researchers go about their work. Some might in fact be public servants working in one of the levels of government, but many will be one step removed and work on commission for those public servants. Whatever the precise relationship, and whether the researcher/analyst works in or for government on commission, the political context of any piece of analysis or research should be recognised and acknowledged.

## 6 Conclusions

To the extent that urban analysis is a form of evaluation and is something done by practitioners of planning, broadly defined, then evaluation research is one of the core practices and competencies of urban planners. Whether they are preparing plans and policies or assessing development proposals in the light of these plans, planners are engaged, unavoidably, in the practice of evaluation. However, just as

M. Jourdain in Moliere's *Le Bourgeois Gentilhomme* did not realise he had been speaking prose all his life, many planners might not realise they have also been practising evaluators. Of course, they might not choose to add this to their CV, but typically it pays to be aware of some of the key principles of evaluation outlined above as this will also, typically, make one a better planner.

Even if professional planners or evaluation researchers are not themselves formally responsible for making decisions about development proposals, strategic planning options or the continuation of programs of intervention, we still tend to hold to the idea(l) that formal decision-makers will pay attention to the professional assessments and evidence presented by planners and evaluators. This is not to say that formal decision-makers—duly elected politicians for example—always accept the advice of their professional planners, evaluators or urban analysts, not least because decisions require the application of political as well as scientific or professional judgement. But if decisions are made that fly in the face of substantial bodies of evidence and professional or scientific advice, then politicians might be held to account and even rejected by their constituents at future elections.

Evaluation research is all about values, goals and objectives—they serve as the benchmarks or yardsticks against which we collect and apply evidence. This is not to say that researchers must accept or agree with these values, simply that we should use these as yardsticks in assessing the success or failure of an intervention or a policy program. We have seen though that success or failure is rarely a simple binary choice. While not impossible, it is unlikely that a policy or intervention is an unequivocal failure in all its dimensions, or indeed a success across the board. Like the curate's egg, it is more likely to be good in parts. Perhaps the greatest challenge facing evaluation researchers lies in avoiding the conclusion of uncertainty; that we cannot say one way or another whether an intervention is successful or not and that further research is necessary. This is sometimes what distinguishes evaluation researchers working as private consultants from those who remain wedded to academic principles: the consultant evaluator will deliver a clear conclusion on time, whereas academic evaluators have been known to take a long time failing to reach a conclusion.

As in many other fields of research, there are craft skills to be developed and applied in evaluation research and practitioners should be looking to continuously improve these. Experience in practising this craft should also bring a heightened awareness of the politics of evaluation research. This might manifest itself in being able to convey conclusions in ways that capture the complexity of the enterprise

without lapsing into ambiguity or inconclusiveness, or that delivers a message of relative failure in ways that do not antagonise the champions of a policy. This is not to say that uncomfortable truths should be avoided, merely that there are ways of speaking truth to power that increase the chance of powerful listening.

Evaluation research is an essential component of urban analysis, broadly defined, and can provide the foundation for better planning and the delivery of better planning outcomes. Like all applied research its practitioners face a number of significant challenges in designing and delivering high-quality evaluation research, including the conceptualisation of what constitutes high-quality work. Because one of the important foundational principles of evaluation research is to speak truth to power, even if the powerful do not always appreciate being told their favourite policy measure is not succeeding in its own terms, it is not always appropriate to judge the quality or value of a piece of evaluation research on the basis of its acceptance—enthusiastic or otherwise—by those commissioning it. But, by understanding and applying its foundational principles and by developing and applying the craft skills of good research, the evaluation practitioner will at least know they have done their best in the challenging environment of evaluation research.

### **Key Points**

- Urban analysis covers many fields, from the assessment of development proposals through research on poverty and inequality in towns and cities to studies of the impact of urban policy measures.
- Evaluation research is an important element of urban policy analysis and involves a range of research techniques and approaches. Few, if any of these research techniques and approaches relate only to urban phenomena or to evaluation work.
- The philosophy, theory and practice of evaluation are subject to constant debate and competing paradigms of evaluation wax and wane. At present, realistic evaluation offers a compelling account of how to design and conduct evaluations that are empirically rigorous and conceptually plausible.
- Evaluation research necessarily involves making judgements about the value or success of something. These judgements are based on criteria, which are typically not determined by evaluators, but by the designers of policies and interventions.
- Making judgemental assessments, even if others have formal responsibility for accepting (or not) and applying them, is part of an inherently political process. Good evaluators recognise this and try to manage it without making unfounded claims to be ‘above politics’.
- The practice of evaluation presents a number of ethical as well as political challenges with the obligation to present truthfully the results of a rigorous evaluation being one of the most important, regardless of the aspirations of those commissioning the evaluation. ‘Speaking truth to power’ remains a fundamental element of good evaluation practice, but remains also a complex principle to apply in practice.



### Further information

For those wanting to get further information about evaluation research, there are a number of useful suggestions for further reading:

- Alan Clarke and Ruth Dawson (1999) *Evaluation Research: An Introduction to Principles, Methods and Practices*, London: Sage
- David Taylor and Susan Balloch (2005) *The Politics of Evaluation: Participation and policy implementation*, Bristol: The Policy Press
- Pawson, R. (2006) *Evidence-based policy: A Realist Perspective*, London: Sage
- Colin Robson (1993) *Real World Research: a Resource for Social Scientists and Practitioner-Researchers*

## References

1. Patton MQ (2020) Evaluation use theory, practice, and future research: reflections on the alkin and king aje series. *Amer J Eval* 41(4):581–602
2. Kemp S (2012) Evaluating interests in social science: beyond objectivist evaluation and the non-judgemental stance. *Sociology* 46(4):664–679
3. Scriven M (1991) Prose and cons about goal-free evaluation. *Eval Pract* 12(1):55–62
4. HY R (1978) Some aspects of symbolic education policy: a research note. In: *The Educational Forum*. Taylor & Francis
5. Boaz A, Davies H (2019) *What works now?: evidence-informed policy and practice*. Policy Press
6. Newman J, Cherney A, Head BW (2016) Do policy makers use academic research? Reexamining the “two communities” theory of research utilization. *Public Adm Rev* 76(1):24–32
7. Shaxson L, Boaz A (2020) Understanding policymakers’ perspectives on evidence use as a mechanism for improving research-policy relationships. *Environ Educ Res*, 1–7
8. Slater T (2006) The eviction of critical perspectives from gentrification research. *Int J Urban Reg Res* 30(4):737–757
9. Harré R, Srivastava N (1983) *An introduction to the logic of the sciences*. Springer
10. Bhaskar R (2013) *A realist theory of science*. Routledge
11. Pawson R, Tilley N (1977) *Realistic evaluation*. Sage
12. Campbell DT (1971) Reforms as experiments. *Urban Affairs Quart* 7(2):133–171
13. Campbell D, Stanley J (1963) *Experimental and quasi-experimental evaluations in social research*. Rand Mc-Nally, Chicago
14. Oakley A (2000) *Experiments in knowing: gender and method in the social sciences* 2000
15. Guba EG, Lincoln YS (1989) *Fourth generation evaluation*. Sage
16. Rossi PH, Freeman HE, Wright SR (1985) *Evaluation: a systematic approach*. Beverly Hills, CA: Sage
17. Chen H-T, Rossi PH (1983) Evaluating with sense: the theory-driven approach. *Evaluat Rev* 7(3):283–302
18. Chen HT (1990) *Theory-driven evaluations*. Sage Publications
19. Robson B, Bradford M (1995) *An evaluation of urban policy*. Paul Chapman Publishing Urban Policy Evaluation, London
20. UN-HABITAT, Monitoring and Evaluating National Urban Policy: A Guide (2020) UNHABITAT: Nairobi

21. Burton P, Goodlad R, Croft J (2006) How would we know what works? Context and complexity in the evaluation of community involvement. *Evaluation* 12(3):294–312
22. Burton P (2009) Conceptual, theoretical and practical issues in measuring the benefits of public participation. *Evaluation* 15(3):263–284
23. Rossi PH, Lipsey MW, Freeman HE (2004) *Evaluation: a systematic approach*. Sage Publications
24. Burton P (2007) Modernising the policy process: making policy research more significant. *Policy Stud* 27(3):173–195
25. UNESCO, Recommendation on science and Scientific Research. 2018, UNESCO: Paris
26. Wildavsky AB (1989) *Speaking truth to power*. Transaction Publishers
27. Robinson EW (2008) *Ancient Greek democracy: readings and sources*. Wiley
28. Head BW (2013) Evidence-based policymaking—speaking truth to power? *Austr J Public Administr* 72(4):397–403