# 7 Designing evaluation strategies

**Matthew James**[1]

**Abstract**

This paper provides an overview of evaluation strategies and their development, with a particular focus on Indigenous policy. The paper starts by outlining the purpose of evaluation and, in doing so, highlights some gaps in the current evidence base. It then looks at the role of evaluation in the policy process and the issues that can arise if evaluation is simply seen as an activity that is separate from policy and that starts after policy is developed. The key message of the paper is that if evaluation is to inform policy it needs to be built into policy processes.

The paper highlights the importance of monitoring and the need to estimate counterfactuals (what would have happened in the absence of a policy or program). It then comments on the importance of involving Indigenous Australians in the evaluation of Indigenous programs before concluding with an overview of some of the practical challenges in undertaking evaluations in Australia.

## 7.1 Why conduct evaluations?

It may seem unusual to ask why evaluations should be conducted, but this is a good question as evaluations, not to mention rigorous evaluations, are not always the norm. The Closing the Gap Clearinghouse, for example, is identifying several areas in which good quantitative studies are lacking. For example, the evidence base from which to increase school attendance in remote Indigenous communities is very thin. Gaps in the evidence base are not unique to Indigenous policy and are not unique to Australia.

If the use of evaluation and good evidence is to be enhanced then it is useful to understand the factors that work against the use of good evidence. One obvious factor is a desire to 'just get on with it' and implement policies and programs. While

[1] Branch Manager, Performance and Evaluation Branch, Department of Families, Housing, Community Services and Indigenous Affairs.

an understandable and in some ways laudable sentiment this sort of thinking is ultimately self-defeating as programs and policies are not developed and implemented for their own sake but rather to improve outcomes — good evaluation and analysis are often needed to assess the impact of policies on outcomes. It is obviously not enough just to see if outcomes have improved but to establish that outcomes have improved as a result of the policy and not some other factor.

Donald Campbell, who is sometimes described as the father of modern evaluation, expressed a well known vision for an experimental society that would be committed to reality testing, to self-criticism and to avoiding self-deception (Oakley 2000). However, Campbell recognised the forces that worked against the use of good evidence in policy formulation. For example, Campbell noted in 1969 that 'specific reforms are advocated as though they were certain to be successful'. He also noted that:

> If the political and administrative system has committed itself in advance to the correctness of efficiency of its reforms, it cannot tolerate learning of failure. To be truly scientific we must be able to experiment. We must be able to advocate without that excess of commitment that blinds us to reality testing. (Campbell 1969)

Willingness to test various approaches is not just an issue for governments and policymakers. Governments need to be allowed to try different approaches and, if they do not work as well as expected, should not face criticism for trying something different. If failure is not allowed, true experimentation is not possible. US President Lyndon Baines Johnson expressed the following frustration about the Great Society programs:

> I wish it had been different. I wish the public had seen the task of ending poverty the same way as they saw the task of getting to the moon, where they accepted mistakes as a part of the scientific process. I wish they had let us experiment with different programs admitting that some were working better than others. It would have made everything easier. But I knew the moment we said out loud that this or that program was a failure, then the wolves who never wanted us to be successful in the first place would be down upon us at once, tearing away at every joint, killing our effort before we even had a chance. (Andrew 1999)

The concern that President Johnson expressed is still relevant today as we do not have all the answers to overcoming disadvantage or to closing the gap on Indigenous disadvantage. Some of the best evidence we do have, such as the impact of high-quality early childhood education, comes from a small number of high-quality evaluations. Without evaluations such as the evaluation of the Perry preschool program much of our evidence would not exist.

There is no foundation to the notion that we have all the answers about overcoming socioeconomic disadvantage and that further evidence is not required. Even a casual

reading of the evidence base confirms this. However, we do know quite a bit and a lack of knowledge should not be used as an excuse for inaction. As an example, if an Indigenous community has too few police — based on need and compared to other communities — there is a good argument that more police should be provided. A detailed study is not required to make this assessment.

A key reason to conduct evaluations is to influence and effect policies, but it is not valid to assess the utility of an evaluation by simply assessing its impact on policy. A high-quality evaluation may have no impact on policy due to a range of factors, including political and other judgements. By their nature evaluations are backward looking and are not designed to tell policymakers what to do next. Working out what to do next requires policy-relevant analysis, not just an evaluation of an existing policy or program.

Simplifying the evidence base is also unwise. Evaluations are important — the detail matters and this detail comes from good evaluations and research. For example, while charter schools in the United States do not, other things being equal, produce better results than public schools, some charter schools are highly successful (Dobbie and Fryer 2009). In addition, the fact that some high-quality early childhood education programs have large positive impacts does not mean that all early childhood education programs have similar impacts, which underscores the importance of rigorously evaluating particular polices and approaches.[2]

Pawson and Tilley advocate what they describe as 'realistic evaluation'. A key focus of this approach is to be careful about universal claims.

> Realist evaluation steers a path between making universal claims about what works, and focusing on the particulars of specific measures in specific places relating to specific stakeholders. Thus it places no faith in black-and-white (or even red, amber and green) policy prognostications of the kind that suppose that street-lighting works to reduce crime, or that mentoring programmes for disaffected youth are harmful, or that 5-fruit-and-veg-portions-a-day health education initiatives have a null effect. (Pawson and Tilley 2004)

[2] In late 2012, a major evaluation of the US early childhood education program for disadvantaged children, HeadStart was released (http://www.acf.hhs.gov/programs/opre/resource/third-grade-follow-up-to-the-head-start-impact-study-final-report-executive). The evaluation, which involved random assignment to a treatment and control group, concluded that, in summary there were initial positive impacts from having access to Head Start, but by the end of third grade there were very few impacts found for either cohort in any of the four domains of cognitive, social-emotional, health and parenting practices. The few impacts that were found did not show a clear pattern of favourable or unfavourable impacts for children.

## 7.2 The need for better evidence — some lessons from the Closing the Gap Clearinghouse

The Closing the Gap Clearinghouse is systematically bringing together the evidence on 'what works' in closing the gap in Indigenous disadvantage. However, the Clearinghouse is constrained by gaps in the evidence base. In many instances, overseas studies have to be cited and, in some areas, good quantitative Australian studies are lacking.

One large issue that has arisen from some of the Clearinghouse products is the general nature of some of the evidence. For example, the evidence base is not informative as it first seems to show that 'bottom-up' approaches work best or that involving local people is key. First, arguing that bottom-up approaches work best provides no advice on what programs or policies to implement. A highly effective program could be implemented using a bottom up approach but so could a program that has consistently been shown in net impact studies to have no effect. Second, how should bottom-up approaches be developed? This is not an easy question to answer as the approach required varies from place to place and the Australian evidence on which to base this question is relatively thin.

Evidence from the Poverty Action Lab is showing that in a development context top-down approaches actually work better for some policies than bottom-up approaches (J-PAL 2012). The details matter. In addition, the authors of a major World Bank publication *Policy Research Report: Localizing Development: Does Participation Work?*, conclude that while involving local people can have positive impacts the benefits on income poverty are modest. A key finding of the report is that care is required with local development projects as, in some instances, projects can be captured by local elites and more disadvantaged community members can be marginalised. The key lesson from the World Bank study is that the evidence base for community development approaches is much thinner than is often thought and, more importantly, community development approaches do not provide a panacea for overcoming disadvantage:

> Evaluations of participatory development efforts improved somewhat between 2007 and 2012, generating some new evidence. However, the evidence base for most questions relevant to policy remains thin, and far too little attention is still paid to monitoring and evaluation. Project design continues to show little appreciation of context, and inflexible institutional rules fail to internalize the complexity inherent in engaging with civic-led development. Unless these problems are addressed, participatory development projects will continue to struggle to make a difference. (Mansuri and Rao 2013)

## 7.3 Evaluation and policy development

The key to a good evaluation strategy is to build evaluation and analysis into policy design. Evaluation should not just be seen as something that happens after a policy is implemented — if evaluation is not considered from the start many types of assessment will not be possible. Evaluation may not be possible at all unless it is built into the policy development budget.

Random control studies are a very obvious example of why evaluation should be considered in policy design. As random control trials can involve providing a treatment to one group and not to a control group, this sort of approach has to be built into policy design, by definition.

Another key reason that evaluation should be built into policy design is that it can be critical to have access to key data before a policy commences if that policy is to be evaluated well.

It is sometimes implied in the Indigenous policy literature that baseline data should be collected before a policy is implemented. While there is an element of truth in these claims they can be misleading and may be counterproductive. The key issue is not whether baseline data are collected but whether the data to make a good assessment are available. If key data are lacking then that data should ideally be collected before a policy commences. However, there is no need to collect existing data into a baseline study before a program or policy commences if that data are readily available and can be produced at any time.

The confusion about the term 'baseline' in an Australian context comes from applying concepts from developing countries that have less relevance in Australia. In developing countries, basic data are often missing, hence the desirability of collecting baseline data before a program or policy commences. On the other hand, Australia has relatively good existing data. For example, data on school attendance are readily available. Insisting that baseline data be collected in an Australian context can be counterproductive and may divert attention from what is really required. If the data are already available and can be reproduced later then that should not be the initial focus. The initial focus should be on collecting any additional data and information that does not already exist. The term baseline does not even appear in one of the best known evaluation textbooks, *Evaluation a Systematic Approach* (Rossi, Lipsey and Freeman 2004), but it is mentioned in

textbooks that discuss evaluation in a development context (Bamberger, Rugh and Mabry 2012).[3]

If new data and information are required to properly assess a policy the collection and timing of that data collection should be built into the initial policy design otherwise key data will either not be collected at all or will come too late to inform policy development and implementation.

Another issue regarding the interface between policy and evaluation is the desirability that policy be based on a clear theory of action. Unfortunately, this is not always the case and, as a result, evaluators sometimes need to develop a theory of action before an evaluation is conducted. Ideally, policy documents should provide a clear logical link between the action and the desired outcome. Otherwise evaluators have to spend time reproducing what should really be a basic feature of policy development. In some instances, as Rossi et al have argued, where program goals and objectives are very unclear or implausible meaningful evaluation may not be possible (Rossi, Lipsey and Freeman 2004, p. 137).

Another issue is being realistic about likely policy outcomes and goals. If policy outcomes and goals are not developed carefully, effective policies may be seen as failures not because they are failures but because the initial goals were unrealistic. As an example, while place-based approaches are popular in Australia, it is important to be realistic about what this sort of approach can achieve. Place-based approaches often have a limited impact on service systems as those service systems are managed and implemented on a whole-of-jurisdiction basis. Given this, it is not realistic to expect place-based approaches to have a large impact on service systems unless those who are tasked with implementing the place-based policy are given the authority to change those service systems.

The recent evaluation of the Groote Eylandt Regional Partnership Agreement showed that this agreement has been successful (MacDonald and Browne 2012). Two key elements of this success are worth noting: taking sufficient time to develop a local plan that all the parties agreed to and agreeing on a realistic set of actions that were then completed.

In his famous article, the 'Iron law of evaluation and other metallic rules', Peter Rossi argued that the expected net value of any net impact assessment of any large scale social program is zero. Rossi cited examples of effective programs and also

3 For example, the Michael Bamberger, Jim Rugh and Linda Mabry book, *RealWorld Evaluation: Working Under Budget, Time, Data and Political Constraints*, which outlines the challenges involved in conducting evaluation in less developed countries, has extensive material on the collection of initial baseline data.

noted that a persistent problem is that policy often underestimates the complexity of the social world and that as a result 'we are overly optimistic about how much an effect even the best of social programs can expect to achieve' (Rossi 1987). These comments, which were written in 1987, are still relevant today. If realistic goals are not set, effective policies may be seen to have failed.

## 7.4 Monitoring

A common expressed frustration with policy evaluations is that they can be too late to inform the ongoing operation of a program or policy and for policy design. One way to avoid this frustration is to ensure that good monitoring is undertaken. Good monitoring can also make any evaluation easier as key data is being collected and analysed throughout the life of the program or policy.

For some programs and policies good monitoring may be all that is required — it may not be necessary to conduct a formal evaluation. However, good monitoring requires not just the collection of data but also good analysis. There is little point in collecting reams of data if that data are not being effectively analysed.

## 7.5 Estimating the counterfactual

Much of the evaluation literature focuses on how to estimate the counterfactual — what would have happened without the policy or the program Often popular or public discussions of policy issues proceed as if we do not need to worry about counterfactuals. Advocates will sometimes cite either improvements to or deteriorations in outcomes and assert that they somehow relate to the value or failure of policy. In reality, outcomes can change due to many factors unrelated to a program or a policy, including existing trends and the state of the economy. While this is an obvious point, it should never be forgotten.

Estimating the counterfactual for an evaluation of an Indigenous policy can sometimes be challenging. For example, for the evaluation of the Northern Territory Emergency Response (NTER) it was difficult to identify comparable communities that could act as a type of control group. The NTER covered such a large number of communities in the NT at once that it was not possible to identify other communities in the NT itself that could have acted as a good control group. It would have been possible to compare trends with Indigenous communities in other parts of Australia but this would have been problematic as the NT is unique in that it is the only jurisdiction in Australia where the vast bulk of the Indigenous population lives in remote communities. For the Cape York Welfare Reform evaluation it has been

possible to compare trends within other remote Indigenous communities in Queensland.

While it was not possible to identify comparison communities for the NTER communities it was possible to look at trends over time thereby using the outcomes in the NTER communities prior to the NTER itself as a control. It was also possible to use natural experiments across the NTER communities to compare trends. For example, trends in recorded crime in communities that gained additional police as part of the NTER were compared with communities where police numbers did not change (AIC 2011).

A key challenge for any evaluation of place-based initiatives in Indigenous communities is the natural noise in data when small numbers of people are involved. This can mean that simply analysing changes in performance indicators may not be very informative. As an example, NAPLAN results for a small school can be very volatile from one year to the next. If in one year five students are in Year 3 and in the next year seven students are in Year 3 it only takes differences across a small number of students to have a quite large impact on the measured result. For this reason, it is sometimes critical to have access to unit record data.

For the evaluation of Cape York Welfare Reform (CYWR) access to unit record data was very important. There has been a large improvement in school attendance in Aurukun in recent years. The task from an evaluation perspective is to explain why this happened. The fact that it happened does not necessarily imply anything about the success of CYWR, as some other factor may have explained the change. Without access to unit record data we would have struggled to establish whether the Family Responsibilities Commission (FRC) had any effect. FaHCSIA ensured that unit record data on school attendance was matched with data on FRC conferences. This allowed us to see whether school attendance for individual students improved immediately after action on the part of the FRC. Without evidence like this it would have been very hard to establish whether the FRC, or some other factor, was affecting school attendance.

## 7.6 Involving Indigenous people

Much of the policy literature on overcoming Indigenous disadvantage focuses on the importance of involving local people in both policy design and policy implementation. However, despite this focus there have been only limited attempts to systematically collect the perspective of Indigenous Australians at a local level using sample surveys.

Indigenous communities are consulted on many issues and, in some cases, they suffer from a consultation burden (AIC 2011, Chapter 2). However, community consultation processes no matter how well they are conducted do not negate the desirability of conducting surveys using standard statistical techniques. Community consultations that involve public meetings may not illicit the views of people who will not speak in front of others because they feel intimidated. This is not an issue that is unique to Indigenous communities. If policy is to genuinely listen to the voice of Indigenous people then those views need to be rigorously collected.

FaHCSIA has now successfully overseen two large scale surveys in Indigenous communities: the Community Safety and Wellbeing Research Study (CSWRS) and the Social Change Survey that has been conducted as part of the evaluation of CYWR. In both instances local researchers were engaged to conduct the surveys and to help determine how particular questions should be asked. Importantly, local researchers were provided with training and support.

By involving local people the CSWRS and the Social Change Survey were more successful than they otherwise would have been. This owed a lot to the skills of the local Indigenous researchers who were able to ask questions in a way that made sense in a local context and who were also able to build trust, thereby gaining the views of people who are not normally asked for their perspectives.

Local surveys can provide rigorous information on both local aspirations and on people's lived experiences. In many instances, these data can be more important than data from other sources. Indeed for some questions such as those on crime victimisation, sample surveys provide better evidence than police data given that much crime is unreported.

## 7.7 Some key challenges in undertaking evaluations in Australia

The single most difficult challenge in undertaking evaluations in Australia, particularly of place-based initiatives, relates to the difficulty in obtaining access to existing data. Bamberger, Rugh and Mabry noted that because evaluation can threaten programs and personnel, some people who are important data sources may take protective measures by limiting or denying access to information (2012, p. 116). The challenges faced in gaining access to data need to be built into evaluation timetables and support should be provided to those undertaking evaluations to try to ensure that data are available in a timely manner. Evaluators, unlike official auditors, cannot demand that data and information be provided. This has particular relevance where an Australian Government agency seeks to obtain

data that relates to State or Territory policies and programs as an Australian Government agency cannot demand or require that key data and information be provided.

Some of the challenges inherent in obtaining access to data relate to rules around national minimum datasets. As an example, the Australian Institute of Health and Welfare (AIHW) holds several national minimum datasets, including data at a local level. Although the AIHW holds the data the AIHW is not free to release these data even to other Australian Government agencies without the agreement of the State and Territory data custodians. Gaining the agreement of those data custodians can be a time consuming process and there can be a lack of consistency. Some jurisdictions are prepared to release data quickly while others may, at times, take months to make a decision and when they do they may try to place quite stringent restrictions on the release of the data. Lengthy delays in attaining access to data can delay the completion of evaluations and may have a serious impact on the quality of the evaluation itself.

In some instances, national minimum datasets do not exist and it is necessary to directly approach individual Australian Government and State Government agencies to access data. This process can work well if it is based on mutual trust and respect; however, this can cause significant delays. Mutual trust and cooperation are key as it is not reasonable to expect an outsider to have a detailed knowledge of the data that an agency may hold.

Some of the practical issues that face evaluators are often not discussed in the literature but these practical issues such as gaining timely access to data can be the key issue. To be fair, the issues often relate to staff shortages in data areas rather than deliberate delays or obfuscation. Often complex datasets require skilled analysts to extract the data. If there are only a small number of analysts, backlogs can quickly develop.

It is not only timely access to data that can delay evaluations but also timely feedback. If feedback from key agencies and stakeholders on draft evaluation reports is not provided according to agreed timeframes, the evaluation can be delayed.

Practical difficulties in undertaking evaluations need to be built into planning processes for evaluations, and, where access to key data is a key factor, agreement to provide key data for evaluations should be built into initial policy design.

## 7.8 Conclusion

This paper has emphasised the importance of careful planning and the importance of building evaluation and monitoring into policy design. If evaluation and monitoring are not built into the policy development process some types of evaluation will not be possible and important data may not be collected.

## References


AIC (Australian Institute of Criminology) 2011, *Northern Territory Emergency Response: Evaluation Report 2011*, Australian Institute of Criminology, Canberra, Chapter 5.

Andrew J.A. 1999, *Lyndon Johnson and the Great Society*, DEE IR, Chicago, Chapter 2.

Bamberger M., Rugh J. and Mabry L. 2012, *RealWorld Evaluation: Working Under Budget, Time, Data and Political Constraints*, Sage Publications.

Campbell D.T. 1969, 'Reforms as experiments', *American Psychologist*, vol. 24, pp. 404-429.

Dobbie W. and Fryer R.G. 2009, 'Are high-quality schools enough to increase achievement among the poor?', *NBER Working Paper No. 15473*, http://www.nber.org/papers/w15473.

J-Pal (Abdul Latif Jameel Poverty Action lab) 2012, *Community Participation*, http://www.povertyactionlab.org/policy-lessons/governance/community-participation, accessed 2012.

MacDonald M. and Browne M. 2012, *Groote Eylandt and Bickerton Island Regional Partnership Agreement: Progress Evaluation, May 2012*, Department of Families, Housing, Community Services and Indigenous Affairs, Canberra, http://www.fahcsia.gov.au/our-responsibilities/indigenous-australians/publications-articles/evaluation-research/groote-eylandt-and-bickerton-island-regional-partnership-agreement-progress-evaluation.

Mansuri G.R. and Rao V 2013, *Localizing Development : Does Participation Work?* World Bank, Washington D.C.

Oakley A. 2000, 'Experiments in Knowing, gender and method in the social science', New York Press, p. 320.

Pawson R. and Tilley N. 2004, 'Realist evaluation', available at http://www.communitymatters.com.au/gpage1.html.

Rossi P.H. 1987, ‘The iron law of evaluation and other metallic rules, *Research in Social Problems and Public Policy*, vol. 4, 1987.

Rossi P.H., Lipsey M.W. and Freeman H.E. 2004, *Evaluation as Systematic Approach*, Sage Publications.