Comments on a Guide to Evaluation under the Indigenous Evaluation Strategy

Don Weatherburn

Background

By way of background I should mention that I have spent more than 30 years evaluating Government programs and policies, including several Indigenous programs. Examples of these can be found on the web-site of the NSW Bureau of Crime Statistics and Research (BOCSAR), which I formerly directed (see https://www.bocsar.nsw.gov.au/). My comments, however, are restricted to the evaluation of programs and policies designed to reduce Indigenous offending and contact with the justice system. I cannot say whether they apply to other domains of Indigenous disadvantage.

Impediments to rigorous evaluation

Data deficits

Most Indigenous policies and programs in the domain of criminal justice are directed at one of three goals:

(1) Reducing Indigenous offending
(2) Reducing Indigenous re-offending
(3) Reducing Indigenous contact with the justice system

In most cases, progress in achieving (1) requires data on persons of interest (POIs) proceeded against by police, broken down by Indigenous status, offence type, month and area (e.g. Local Area Command/Local Government Area). All jurisdictions have the capacity to provide this data but few actually make it available to researchers outside of Government

Most rigorous evaluations directed at reducing re-offending measure re-offending via re-arrest or re-conviction. To measure progress in reducing Indigenous re-offending requires (at a minimum) access to de-identified unit record file data on persons arrested or convicted, along with a host of other characteristics, including their Indigenous status, their age, their gender and their prior criminal record (see, for example, Yeong and Moore 2020). The Reoffending Database (ROD) developed by the NSW Bureau of Crime Statistics and Research is the ideal set up but most jurisdictions do not have a similar database or, if they do, do not allow external researchers access to it.

The measures most often needed in relation to (3) are (a) the percentage of Indigenous defendants diverted from court, (b) the percentage of Indigenous defendants refused bail, (c) the percentage of Indigenous defendants given a custodial sentence and (d) the number of Indigenous prisoners (adult

or juvenile). The Australian Bureau of Statistics can provide data on (d) (though not at unit record level) but access to data on (a) to (c) is limited and patchy.

One of the main reasons for these data deficits is the rather balkanised nature of most criminal justice systems. Police collect the data they believe they need for law enforcement, courts collect the data they believe they need to manage courts and corrective services collect the data they believe they need to manage custodial and correctional orders. Although efforts have been made to break down these silos, little progress outside NSW has been made in creating data systems that could answer policy questions about the effects and effectiveness policies and programs in reducing re-offending or contact with the criminal justice system, whether by Indigenous or non-Indigenous defendants.

Managerial impediments

Commitment to rigorous and objective policy and program evaluation starts at the top. If the CEO of an organisation and his or her senior offices are not committed rigorous evaluation and its corollary (preparedness to accept disappointing results), then no amount of commitment at lower levels of the organisation will make it happen. In my experience, Ministers and CEOs of justice and police organisations have very little understanding of evaluation, which they often confuse with monitoring.

The problem is not that they lack statistical training. What they lack is (a) an understanding of the central aim of outcome evaluation; that is, to create a counterfactual that will provide insight into what would have happened in the absence of the policy or program and (b) the main strategies available for creating a counterfactual. Most will have heard of a randomised controlled trial but would be instinctively opposed to conducting such a trial. Few would have heard of instrumental variable methods, difference in difference or regression discontinuity.

This ignorance has two adverse effects on the scope for rigorous evaluation. Firstly, CEOs frequently make decisions on the scale of a pilot, its duration, its location and its eligibility criteria that make it impossible to create a counterfactual or that result in a sample size too small to pick up worthwhile effects. These problems could be avoided if early consultation occurred with the evaluation team but in most circumstances decisions about the scale of a pilot, its duration, its location and its eligibility criteria are made before an evaluation team has been assembled. If the CEO and other senior managers had a basic understanding of evaluation principles and strategies, this problem could be avoided.

Secondly, because senior officers often have a poor understanding of outcome evaluation strategies, they often fail to establish the data systems needed to make rigorous evaluation possible. There is a tendency to collect large quantities of information about those who participate in a program or are the target of a policy and nothing or very little about those who could have participated but didn't or who could have been the target of the policy but weren't. In my experience, for example, correctional data systems rarely distinguish between people not accepted onto a program because they were not eligible, and people eligible for a program but not accepted onto it for reasons unrelated to the outcome of interest (e.g. shortage of program places). This makes it impossible to bring instrumental variable methods to bear on the evaluation.

Infrastructure impediments

The time required to establish data systems and data sharing protocols often results in long delays in the completion of outcome evaluations. These delays are often exacerbated by staff who see any change in their practices as an unwarranted additional burden. One solution to this problem is to create centres of excellence, that is, locations where new policies and programs are trialled. Justice and police organisations, however, have also been slow to establish centres of excellence in evaluation that could become the focal point for trialling new programs and policies. An ideal arrangement would be a designated model court and model local area command, staffed by people with training in program management and evaluation and equipped with the data systems necessary to carry out rigorous evaluation of policing, legislative and court reform initiatives.

The creation of a model court and/or model police command would assist in solving another problem in public policy evaluation. Unlike their counterparts in the private sector, Government agencies rarely have any commitment to research and development (R&D). There is a tendency to think that if a program or policy does not work then the program or policy should be abandoned. In fact programs and policies often face teething problems at the beginning that limit their effectiveness. If a process evaluation has been conducted (more on this below), it is often possible to make changes to a program or policy to render it more effective. This entails an ongoing commitment to R&D. The establishment of a dedicated test site for new crime control or new reoffending/contact with the criminal justice system reduction programs would facilitate the transition to a R&D approach to public policy evaluation.

Building evaluation into policy and program design

The draft report is right to observe that evaluation is most effective when it is planned for early and when it is integrated into each stage of policy and program design — from setting policy objectives through to using evaluation findings to inform future policies and programs. It is also critical that Aboriginal and Torres Strait Islander people, perspectives, priorities and knowledges be carefully considered at each of these stages.

There are, however, two elements to this. One concerns Indigenous involvement in policy and program development. The other is Indigenous involvement in evaluation. The first is much easier to achieve than the second. Indigenous involvement in policy or program design can be achieved through a suitably structured form of consultation or through a suitably constituted advisory group or groups. Evaluation design requires certain technical skills that, at present, are in very short supply amongst Indigenous university graduates. There is a pressing need to increase the number of Indigenous Australians with tertiary level training in evaluation research.

Drawing on evaluation evidence when designing a policy or program

Policy makers should examine the existing evidence base at the earliest stages of policy and program design to see what lessons there are from similar policies or programs and this includes assessing the strengths (and weaknesses) of the evidence supporting the assumptions upon which a policy or program is based.

The assumption here, however, is that the notion of 'evidence base' is unproblematic and that all reasonable people would agree on what constitutes 'strong' or 'weak' evidence. It is a sad fact that what most of us trained in scientific method would regard as strong (or weak) evidence is not shared by everyone. I think the Productivity Commission should nail its colours to the mast here and refer the reader to examples of systematic reviews (e.g. the Cochrane and Campbell Collaborations) and explain what it means by strong evidence. A good example (although by no means the only one) is the hierarchy of evidence established in the Report to the US Congress on Evidence-Based Crime Prevention by Sherman et al. (2000). I would only add that I think all outcome evaluations should be accompanied by a process evaluation and that the central goal of process evaluation should to gauge program fidelity, that is, the extent

to which a program was implemented as intended. This is where qualitative research becomes critically important.

An Indigenous Evaluation Threshold Assessment

The draft suggests that an Indigenous Evaluation Threshold Assessment should be carried out. I don't see the need for a threshold. The NSW Government, for example, is issued a directive that no new program (whether directed at Indigenous Australians or otherwise) will receive Treasury funding unless it has an approved evaluation plan. I can't imagine the circumstances in which a justice agency could justify implementing a program designed to reduce Indigenous crime, reoffending or contact with the criminal justice system without evaluating it. The moment you establish a threshold for policy or program evaluation, politicians and bureaucrats will start gaming it. The proposed multi-criteria assessment to determine evaluation priorities is nicely rational in theory but it wide open to manipulation by Ministers who want to spend money on their pet program without the burden of evaluation.

Small scale evaluations of main-stream programs

I understand the motive for suggesting small scale evaluations of large-scale programs but there are risks in evaluating large scale programs in small scale studies.

The obvious one is statistical power. You need a sample size of about 400 in the treatment and comparison groups to have an 80 per cent chance of picking up a 10 per cent difference in re-offending. The risk associated with small scale evaluations with outcomes that are comparatively rare is that you inflate the Type 2 error rate. Before deciding to conduct a pilot evaluation of a program people need to think how big a sample they need to have a reasonable chance of finding an effect if there is one. They also need to think how small an effect has to be before it is not worth having.

The importance of process evaluation

Programs and policies can fail to produce their intended effects for one of two reasons. One is that the assumptions underpinning the program or policy are incorrect. The other, much more common reason, is that the program or policy has not been implemented as intended. The best way to determine whether the latter is the case is to conduct a process evaluation. This is where qualitative methods have a key role to play. The ideal process evaluation is one in which the researcher, in

consultation with the program/policy manager, identifies the conditions that have to be met for the policy to have been properly implemented. The process evaluator then designs a study to assess how well the program/policy has been implemented using these conditions as a guide. Process evaluations are expensive because they are usually labour intensive. However if a program or policy fails to produce its intended effect and you do not know whether it was properly implemented, you cannot tell whether the program or the program implementation process was at fault.

Who should do the evaluation?

Most Government agencies in my experience are poorly equipped to conduct rigorous and objective evaluations. Their research units are often embedded in their policy units, the culture is one of pleasing the Minister and senior officers tend to be highly risk averse. When an evaluation is necessary, the response to these conditions is often to call in outside private sector consultants, often big accounting firms. That way the report if it is controversial can be kept away from public view and senior officers do not have to wear the burden of bringing bad news to Government.

Some private sector agencies do conduct rigorous and objective evaluations of Government policy. Others charge a great deal of money to serve up a slightly modified version of what the Government of the day or the Department Head wanted to hear. There are three ways of avoiding this problem. The first is to ensure that the person letting the tender for the evaluation has a strong grasp of evaluation methods. The second is to establish a suitably resourced evaluation unit within the relevant Government agencies. If this course is chosen, the unit should be kept separate from the policy/program development units in the agency and should report to a different officer (preferably the CEO). The third option is to forge an ongoing relationship with a suitable department in the university sector. The relationship in this case needs to be ongoing (although not necessarily permanent) because it takes time to build up the requisite level of expertise. It is possible to pursue all three of these options, depending on the project at hand.

Steps in undertaking an evaluation

Finally, for what it's worth, here are the steps I think should be undertaken in conducting an evaluation of a Government policy or program

- Identify the objectives of the program and any possible unintended consequences
- Identify key stakeholders

- Determine whether a credible outcome evaluation is feasible

- Assuming it is, formulate a research design

- Consult with key stakeholders to explain the design and receive feedback

- Make any necessary changes to the design and get stakeholder signoff

- Collect data and carry out the analyses

- Present preliminary findings to stakeholders and get feedback

- Carry out any additional analyses required and draft evaluation report

- Circulate draft report to stakeholders for comment

- Revise draft report and send it out to independent assessors for peer review

- Make any necessary changes and finalise report.