# 5 The case for making public policy evaluations public

**Deborah A Cobb-Clark**[1]

**Abstract**

This paper sets out the case for making public policy evaluations public. It first reviews the various challenges associated with impact evaluations, paying particular attention to the unique hurdles involved in evaluating Indigenous policy. Lessons learned from clinical trials registries in medical research are then used to argue that Australian economic and social policy evaluations could be improved by making them public.

## 5.1 Introduction

Efficient and effective public policy must be informed by solid evidence about what actually works, for whom, under what circumstances, and at what cost. Program evaluation plays a critical role in building the evidence base necessary to answer these questions and in providing all levels of government with the necessary information to develop initiatives that allow more to be achieved with the same or, perhaps, even fewer, resources.

This paper sets out the case for making public policy evaluations public. I begin by first reviewing the various methodological, data, administrative, and political challenges that undermine our ability to use program evaluations as a tool for improving public policy decision-making. My focus is strictly on impact evaluations and I pay particular attention to the unique hurdles involved in evaluating Indigenous policy. I then draw on the lessons learned from clinical trials registries in medical research to argue that Australian economic and social policy could be improved by making the evaluation of those policies publically accessible to service providers, other government agencies, researchers, and taxpayers.

1 Director and Ronald Henderson Professor, Melbourne Institute of Applied Economic and Social Research, University of Melbourne; and Institute for the Study of Labor (IZA)

## 5.2 Why are evaluations often not informative?

In theory, program evaluation is very straightforward. One simply randomly assigns some individuals to participate in a particular program (or to receive a specific service) and others to randomly miss out. The former group are then 'treated', while the latter group become the 'controls'. The impact of the program is estimated by simply taking the difference in the outcomes achieved by treated individuals and those achieved by the controls. Matters become somewhat more complicated when individuals cannot literally be randomly assigned to government policy initiatives. However, it is often possible to conduct a credible impact evaluation with a quasi-experimental design utilising exogenous variation (for example, across time, age, locations) in program eligibility or implementation. Random (or plausibly exogenous) variation in the receipt of treatment eliminates the need for complicated econometric techniques to generate estimates of treatment effects.

The reality of program evaluation, however, almost always deviates wildly from the theory. What is straightforward when discussed in the pages of a textbook, becomes anything but when discussed in a policymaker's office. Real world public policy evaluations are conducted under a number of constraints, both methodological and political.

### The methodological constraints: The reality is more challenging than the theory

Some things are simply harder to evaluate than others. Evaluation always becomes harder when (i) a large segment of the population is affected; (ii) the policy is complex; (iii) program implementation or delivery lies in the hands of others; and (iv) individuals have control over their treatment status.

In particular, program evaluation becomes harder as the affected population becomes larger, for two reasons. First, it becomes much more difficult to find a sensible counterfactual or control group. In the United States and Canada, many welfare policy reforms, for example, take place at state or provincial levels, which allows otherwise similar jurisdictions to act as counterfactuals. In contrast, Australian income-support policy is under the purview of the Commonwealth Government and policy reforms tend to affect the nation as a whole. Although it is still possible to use before–after research designs to evaluate Australian income-support policy, much stronger assumptions are needed for identification than would be the case if both geographic and time variations were available (Cobb-Clark and Crossley 2003; Meyer 1995). Second, it becomes much more difficult to ignore the general equilibrium effect of policy reform. Although it might be reasonable to

ignore the way that a small, localised training program affects employment in the Australian labour market as a whole, this assumption is surely less justifiable when the program is implemented nationally.

Program evaluations also become more challenging as the complexity of the underlying policy intervention increases. Particularly challenging social problems, for example, are likely to be met with multifaceted policy initiatives with multiple objectives. Sometimes these objectives are clear. Frequently, however, the policy goals are ill-defined, unarticulated, or even directly contradictory. Moreover, evaluations are always constrained by our ability to actually measure the outcomes that we care about. While it is perfectly legitimate for policymakers to be concerned with things like the extent to which a program engenders a sense of community spirit or empowerment, evaluating it from this perspective requires that we can actually measure these outcomes. Finally, it is not always possible to separately evaluate the individual components of a complex initiative despite policymakers' considerable interest in doing so.

Evaluation can also be tricky whenever the program being evaluated is implemented or administered externally. In this case, it is often necessary for components of the evaluation (for example, recruiting participants, collecting data) to be outsourced to the agencies delivering treatment. It can be very difficult to maintain random assignment in this situation, in part because program administrators may be more accustomed to using their professional judgement to match clients to programs. Unfortunately, this makes it impossible to get an estimate of the program impact *per se* independent of the selection process used to match individuals to that program. At the same time, it is often still possible to get an estimate of the effect of the combined treatment; that is, the process used to assign treatment plus the program itself. In many cases, this will continue to be very valuable since very few real world initiatives are actually directed at randomly selected individuals. Hence an estimate of the combined treatment effect is likely to be of great use to policy makers.

The most challenging — and consequently least plausible — evaluations occur when individuals can influence whether or not they receive treatment. For example, workers may reduce their hours of work in order to be eligible for a government training program. Schools may select certain students over others in order to receiving funding that is dependent on student demographics. In most cases, the success of public policy initiatives relies on individuals and institutions responding to exactly these sorts of economic incentives. However, these same responses can wreak havoc on an evaluation strategy. The problem is quite simple: if treatment status is not randomly (or exogenously) assigned, those in the treatment group will differ from those in the control group in unobservable ways (for example,

motivation or ability) that are potentially related to their outcomes (for example, wage rates or test scores). This means that a simple comparison of outcomes for those who do and do not receive treatment will not necessarily tell us very much about the impact of the treatment itself. There is a raft of non-experimental econometric techniques, including propensity score matching and instrumental variables, that get hauled out in such situations. But at the end of the day, these approaches only deal with that part of selection into treatment that is based on individuals' (or schools') observable characteristics. It is still necessary to rely on a maintained (i.e. untestable) assumption that, conditional on these characteristics, there are no unobservable differences between the treatment and control groups which would affect the outcome we are interested in. This assumption is clearly easier to defend the more data we have and the more characteristics we can take into account.

The final point to make here is that, in the end, program evaluations always rest on the available data. It is simply not possible to evaluate what we cannot observe. It is not uncommon for data limitations to constrain the evaluation questions, the evaluation method, the quality of the evaluation, and indeed whether an evaluation is even possible. It is also important to note that while non-experimental evaluation approaches can be very useful in providing critical information in less than ideal evaluation situations, they are very data intensive relative to experimental and quasi-experimental approaches. One of the most important investments we can make is in data sources which can be used to support public policy evaluation.

## The political constraints: Better than nothing is not the same as good enough

In addition to the methodological constraints described above, program evaluations are typically also subject to a number of time, budget, administrative, and political constraints — which for convenience I will simply label as 'political' constraints. These constraints come in a myriad of forms and have a critical — usually unfortunate — role in shaping the overall evaluation methodology. Increasingly, practical advice in managing these constraints is being sought by researchers engaged in real world program evaluation. For example, Bamberger *et al*. (2004, p. 5) write in their recent article:

> This paper discusses two common scenarios where evaluators must conduct impact evaluations when working under budget, time or data constraints. Under the first scenario the evaluator is not called in until the project is already well advanced, and there is a tight deadline for completing the evaluation, frequently combined with a limited budget and without access to baseline data. Under the second scenario the evaluator is called in early, but for budget, political or methodological reasons it is not

possible to collect baseline data on a control group and sometimes not even on the project population.

The authors go on to make the obvious point that as a result of these constraints, many of the basic principles of program evaluation get sacrificed. Their goal is to provide practical workarounds to yield the best possible evaluation under the circumstances.

What is particularly striking about the Bamberger *et al*. (2004) paper is their realistic portrayal of the situation that most program evaluators find themselves in. Many — perhaps even all — public policy evaluations in Australia are conducted under exactly these sorts of constraints. However, while it may be possible to 'rescue' some semblance of an evaluation strategy with very clever lateral thinking, it is critical to recognise that in the end we may not have actually learned very much. Often 'better than nothing' passes for 'good enough', leaving us as uninformed as ever, despite having spent millions (or tens of millions) of dollars on the evaluation exercise.

## The particular challenges in Indigenous program evaluation

There are unique methodological and political challenges in evaluating Indigenous programs, which I outline here.

First, Indigenous Australians make up only around 2 per cent of the total Australian population and Indigenous communities themselves are often quite small. As a result, many data sources are unsuitable for Indigenous program evaluation because they do not have sufficient numbers of Indigenous respondents for analysis. Even when quantitative analysis is possible, small sample sizes can drastically limit statistical power. This means that, given realistic sample sizes, only very large program impacts are likely to be detected at standard statistical levels.

Second, for cultural, historical, and political reasons it is often argued that Indigenous communities are unique and therefore cannot be meaningfully compared to one another. To the extent that this is true — or we accept it out of cultural sensitivity — it becomes nearly impossible to define a meaningful control group against which to measure impacts.

Third, many Indigenous policy initiatives are targeted at communities. Moreover, the Indigenous population is characterised by fluid, extended family structures and cultural norms for resource sharing. Together these imply that it is very difficult to estimate the effect of treatment on the individuals treated (i.e. a treatment on the treated impact). For example, even though income management theoretically

applies to an individual benefit recipient, in reality it is likely to have substantial spill-over effects on his or her extended family and other community members. As a result, in most cases, we will be estimating parameters which are closer to a community-level impact of the intention to treat.

Fourth, and related to the above, because Indigenous programs are often community-based interventions they need the approval and support of community elders. There is almost no sense in which Indigenous communities are randomly selected for treatment. The effects of the selection process itself — normally long, drawn-out negotiations between government and Indigenous elders — will be a component of what is measured in the estimated treatment impact. It is impossible to identify the effect of the program itself in isolation from these selection effects. That is not a particular problem given that it is unlikely that programs will ever be randomly assigned to Indigenous communities. However, it does complicate our interpretation of the estimated impacts and must be borne in mind.

Fifth, Indigenous policy is often highly political and involves a cast of thousands, including Commonwealth, State, and local governments; social service agencies and non-government organisations; Indigenous representatives and their communities; and a raft of advisors, advocates, and analysts. At any one time, there is likely to be a myriad of interventions affecting the Indigenous population. This means that it is very difficult to evaluate any single program in a particular Indigenous community because a multitude of programs are being delivered simultaneously. If another Indigenous community is used as the counterfactual, it is certainly the case that the 'control' group is also treated — just with a different set of policies and programs. Therefore, standard evaluation techniques provide only an estimate of the marginal difference between one set of interventions and another set, many (indeed most) of which overlap. This is almost never the estimate we want, and in some cases, may not be interesting at all.

Has a lack of Indigenous-specific evaluation limited our ability to learn from past policies? It is impossible for me to say for sure, but it seems exceedingly hard to believe that this is not the case. If nothing else, the continuing gap in Indigenous versus non-Indigenous outcomes in the face of the very substantial resources committed to Indigenous policy clearly indicates that we must do better at finding effective policies that will truly improve the wellbeing of Indigenous Australians. Program evaluation that is well done, methodologically sound, and corresponds to accepted scientific principles is critical to achieving that goal.

## 5.3 The case for making public policy evaluations public

### The example of health care evaluation

Ten years ago the *British Medical Journal* published an editorial arguing for increased transparency in economic evaluations of health care as a means of ensuring higher methodological quality. Specifically, the authors argued that:

> We need periodic methodological assessments of economic evaluations using adequate sampling frames. The assessments should be ongoing and publically accessible. Unless swift action is taken, low methodological quality risks bringing the practice of economic evaluation into disrepute — an outcome that is in no one's interest (Jefferson and Demicheli 2002).

This increasing pressure for greater transparency in health care evaluations resulted in part from several systematic reviews conducted in the early 1990s which cast doubt on the scientific reliability of published evaluations. Each of these reviews argued for improvements in the standards for conducting and reporting economic evaluations (see Jefferson and Demicheli 2002). In short, increased transparency and wider dissemination of results were viewed as fundamental to raising the methodological quality of economic evaluations in health care.

The reasons for this are not hard to understand. Despite the widespread use of randomised control trials (often regarded as the 'gold-standard') in health care interventions, it is often the case that results are not widely disseminated. Indeed, many experts may never learn that a trial has taken place. Gold and Studdert (2005) point to a number of ways that incomplete, non-systematic reporting of results undermines the randomised control trial methodology in health care research. First, the results of many trials are never published and those that ultimately are published are systematically different from those that are not. Specifically, studies that show the efficacy of the intervention are simply more likely to be published. This sort of positive publication bias makes it impossible to form valid judgements about an intervention's true effectiveness from the published literature. Second, there may be strong financial incentives to withhold negative results and suppress data.[2] In particular, Gold and Studdert (2005) point to the recent legal case against the pharmaceutical company GlaxoSmithKline (GSK) which manufactures the popular anti-depressant Paxil. Although not officially approved for children, millions of Paxil prescriptions were nonetheless written for children. The legal case revolved

[2] The authors refer to the first as a form of scientific misconduct and argue that the second may constitute fraud.

around GSK's failure to acknowledge and report the results of several studies that had raised doubts about Paxil's effectiveness and safety for children. The plaintiff argued that GSK had a duty to disclose negative studies, not just positive ones.

One important response to the call for greater transparency in health care evaluations has been the establishment of clinical trials registries. In particular, in 2005 the Australian New Zealand Clinical Trials Registry (ANZCTR) was established at the University of Sydney as part of the World Health Organization Registry Network. It accepts trials for registration from all countries around the world and from the full range of therapeutic areas, including trials of pharmaceuticals, surgical procedures, preventative and lifestyle measures, and rehabilitation strategies.[3] Registration of trials occurs before the first patients are recruited. The ANZCTR is overseen by an advisory board and a substantial amount of initial funding was provided by the Australian Government through the National Health and Medical Research Council (NHMRC).

The registration of Australian health care trials with an institution such as ANZCTR is voluntary. While some argue that mandatory registration risks manufacturers' proprietary information and undermines incentives to engage in research and development, others argue that those risks must be balanced against the benefits of registration, which only occur if registration is comprehensive (Gold and Studdert 2005). In practice, however, registration has become *de facto* mandatory for those seeking to publish the results of their trials. Since early 2004, the International Committee of Medical Journal Editors (ICMJE) has made trial registration a necessary condition for the publication of any manuscript reporting trial results (Gold and Studdert 2005). In the United States, it is a legal requirement that many types of medical trials be registered.[4]

## Economic and social policy evaluation

There are many parallels to be drawn between evaluations in health care and program evaluation in economic and social policy more generally. Most importantly, increased transparency and wider dissemination of results are absolutely essential to improving the quality and information content of our economic evaluations of Indigenous policy, education initiatives, and income-support, disability, and job training programs etc. Moreover, the arguments in favour of an institutional arrangement like a clinical trials registry are as compelling in these areas as they are in the area of health care.

[3] See http://www.anzctr.org.au.

[4] See http://www.clinicaltrials.gov.

First, initiatives in Indigenous, education, or income-support policies can have as profound an effect on individuals' lives as those in health care. It follows then that it must be as important to do a credible job of evaluating them. Second, although meta-analyses of the program evaluation literature are nearly nonexistent, it must surely be the case that — as in health care — positive publication bias skews the published results. Here, however, 'positive' often refers not to the efficacy of the particular drug or treatment, but rather to the desirability of the program from bureaucrats' or politicians' perspectives. Third, the Paxil case illustrates the tension between private (manufacturer) and public interest in publicising the results of medical trials. A similar tension arises when government departments or non-governmental organisations have a private incentive to withhold information about the impact of particular programs or policy initiatives.

In short, many of the factors which led to the call for greater transparency in health care evaluations a decade ago are relevant in economic and social program evaluation today. We must raise the standards of program evaluation. Greater transparency and wider dissemination of results are central to achieving these goals. In particular, greater transparency would (i) put pressure on evaluators to lift their game; (ii) allow evaluations themselves to be evaluated against sound scientific principles so that we can make judgements about which to weight more heavily and which to ignore; (iii) provide an opportunity for truly informed public debate about the issues facing us; and (iv) substantially enhance our chances for sound decision-making.

At the same time, the comparison with health care trials is not perfect. Publication has always been more critical to the private interests of pharmaceutical and medical device manufacturers, thus the position taken by the ICMJE in supporting trial registration has had a significant role in ensuring that trials do in fact get registered. The same cannot be said of economic and social policy evaluations more generally. In fact, governments often work hard to ensure that results are not made publically available. Moreover, randomised control trials are less common outside the health care arena. Evaluation in other policy areas relies more heavily on quasi-experimental (that is, 'natural') and non-experimental evidence. This implies that methodologies are much more complex, which would require much more flexible reporting systems. Finally, several key drug failures (for example, Paxil) have focused the collective mind on the importance of sound health care evaluations in a way that is unlikely to happen in other policy arenas.

These caveats imply that rather than adopting the existing medical trials registries as is, the basic principles underlying them will need to serve as a framework for developing unique institutional arrangements that can achieve the same objectives

in other policy areas. They do not, however, strike me as arguments against making public policy evaluations public.

## 5.4 Conclusion

At this moment, the Australian Government is poised to spend literally billions of taxpayer dollars on major social initiatives in Indigenous policy, educational reform policy, and supporting the disabled policy. This is an enormous commitment of public resources which comes with huge opportunity costs given the political imperative to return the government budget to surplus. Sound, independent program evaluation will be crucial to ensuring that we receive value for money.

Unfortunately, our current evaluation system generally produces poor-quality evaluations that in the end do not tell us very much. Often evaluations are conducted within the very government agencies responsible for meeting program objectives. When external evaluators are used, it is common for the government to insist that the results not be published. In short, the results of these evaluations are typically not independent, transparent or widely distributed. As a result, methodological quality is undermined. All of this is inconsistent with the move to evidence-based policy and undermines our ability to deliver on closing the Indigenous gap, raising educational achievement, and reducing social exclusion.

Public registration of economic and social program evaluations will not completely resolve these problems, of course, but is an important step in the right direction. In addition, we need to work harder to ensure that a sensible evaluation plan is embedded — and funded — in the design of the program from the start. In particular, capacity constraints imply that welfare reform in Indigenous communities, educational reform, and the national disability insurance scheme will be rolled out over time in certain locations or for certain groups of individuals. These rollouts — if planned properly now — will allow high-quality program evaluations of these initiatives to take place.

It is also critical that we have systematic program evaluations that are truly independent of government. The lack of a willingness to commit to eventual publication of results has meant that Australian academics are increasingly disengaged from evaluations of major economic and social initiatives. This is unfortunate because there is a great deal of evaluation expertise within the academic community; moreover, the academic publication process has a critical role to play in quality assurance. One potential mechanism for supporting this would be a separate agency which commissions all policy evaluations on behalf of the government, but which is independent of government (like the Reserve Bank of Australia or the

Productivity Commission). All evaluations conducted by (or commissioned through) this agency could then be published externally, perhaps with a short embargo period.

Finally, all components of any program evaluation, including the unit-record data on which it rests, must be widely and publically available, so that results can be replicated and confirmed. Widespread publication of evaluation results must become the norm.

If we truly wish to make progress on the economic and social agenda we have set ourselves, better than nothing will not be good enough.

## References


Bamberger, M., Rugh, J., Church, M, and Fort, L. 2004, 'Shoestring evaluations: designing evaluations under budget, time and data constraints', *American Journal of Evaluation,* 25(1), Spring, pp. 5–37.

Cobb-Clark, D.A. and Crossley, T. 2003, 'Econometrics for evaluations: an introduction to recent developments', *Economic Record,* 79(247), December, pp. 493–513.

Gold, J.L. and Studdert, D.M. 2005, 'Clinical trials registries: a reform that is past due', *The Journal of Law, Medicine, and Ethics,* 33(4), December, pp. 811–20.

Jefferson, T. and Demicheli, V. 2002, 'Quality of economic evaluations in health Care: it is time for action to ensure higher methodological quality', *British Medical Journal,* 324, 9 February, pp. 313–14.

Meyer, B.D. 1995, 'Natural and quasi-experimental experiments in economics', *Journal of Business and Economic Statistics*, 13(2), April, pp. 151–61.