CRIME AND JUSTICE
Bulletin

NSW Bureau of Crime
Statistics and Research

Contemporary Issues in Crime and Justice Number 133

July 2009

Policy and program evaluation:

Recommendations for criminal justice policy analysts and advisors

Don Weatherburn

INTRODUCTION

Over the last 20 years in Australia the attitude of Government to policy evaluation has changed dramatically. Ministers (and Treasury officials) are no longer content to rely on anecdotes from agency heads as evidence that their programs are working. Rigorous evaluation is usually required, especially if the program is controversial or expensive. This is an encouraging development because it provides greater assurance that taxpayer funds are being wisely spent. Many of those responsible for fashioning or implementing new policies, however, though experts in their own field, have only a limited understanding of the circumstances in which rigorous evaluation is possible.

Failure to understand the preconditions for good evaluation can lead to frustration and disappointment on the part of evaluators and clients alike. Evaluators become frustrated when they are called in to evaluate a program after key decisions have already been taken about where the program will be tested, when it will commence, how much money will be provided for the evaluation, what the evaluation will address and/or when the evaluation report will be complete. Policy advisors and program managers, on the other hand, become irritated when an evaluator wants to sit in on meetings to design a program (the 'tail' wagging the 'dog'), when an evaluator says it is impossible to tell if a program is achieving its intended outcomes or when he or she makes what seem to be unreasonable demands in relation to the scale, start date and location of a pilot program.

Few policy advisors have the time to become experts in the research techniques and methods employed in policy evaluation. Fortunately, one does not need to be an expert on policy evaluation to know when an evaluation is possible and when it is not. An understanding of basic principles is sufficient to avoid making costly and/or embarrassing mistakes. With this in mind the present bulletin presents and explains four key principles that anyone thinking about commissioning or requesting an evaluation in criminal justice or crime prevention should know.

SOME PRELIMINARY CONSIDERATIONS

In some contexts it is important to distinguish 'policies', 'programs', 'initiatives', 'interventions' and 'operations' etc. In this bulletin we will make no distinction between these things and speak generally about programs. A program for our purposes is simply a set of actions undertaken to reduce crime or make the criminal justice system more equitable, effective and/or efficient. So defined, the term includes everything from police operations to increase the arrest rate of persistent burglars though to policies designed to reduce the number of cases where defendants change plea on the day of their trial, through to rules designed to ensure that legal aid is available to people who cannot afford to pay for legal representation themselves.

TYPES OF PROGRAM EVALUATION

There are two main types of evaluation: outcome evaluations and process evaluations. The aim in an outcome evaluation is to see whether a program is producing its intended outcomes and/or any unintended outcomes. The aim in a process evaluation is to see whether a program is operating as planned. When programs fail to achieve their intended outcomes, process evaluations help us understand why. They also help in identifying ways of making programs more efficient and effective. The rest of this bulletin, nonetheless, is about outcome evaluation.

PRINCIPLE 1: ALL EVALUATIONS REQUIRE WELL-DEFINED AND MEASURABLE AIMS

If the aims of a program are not clear, its intended outcomes cannot be measured and the program cannot be evaluated. This might seem like commonsense but the aims of some programs are troublingly vague.

This bulletin has been independently peer reviewed.

The aim of one Australian restorative justice program, for example, is to increase offender's awareness of the consequences of their actions and reintegrate them back into the community. This aim sounds clear until one asks how its achievement might be measured. How, for example, do we measure an offender's 'awareness of the consequences of their actions'? How do we tell when an offender has been 'reintegrated' back into the community? Some might regard an offender as reintegrated if he or she is not convicted of a further offence for at least two years, no matter how antisocial in other respects they may be. Others might require more — a job, for instance, or remorse, or evidence of commitment to conventional social norms. What appears to be a successful program on one of these interpretations could quite reasonably be judged an abject failure on others.

Ideally, the intended outcomes of the program should be stated in quantitative terms. It is better to say, for example, that the aim of a program is to reduce crime by (say) 10 per cent than simply to say that the aim is to reduce crime. If you have no idea what reduction in crime or recidivism (or any other outcome) to expect, it is better to state the minimum reduction necessary to make the program worth the money and effort being spent on it than to say nothing at all. The reason for this is explained in connection with Principle 3 below.

PRINCIPLE 2: ALL EVALUATIONS REQUIRE A BASELINE OR A CONTROL GROUP (OR BOTH)

All outcome evaluations seek to generate a counterfactual—that is, to create a situation where we can tell what would have happened to a particular outcome (e.g. crime, recidivism, court delay) had the program not been introduced. There are numerous ways of doing this but they all involve constructing a control group or baseline or both. In this section we highlight the main ways in which control groups and baselines are used.

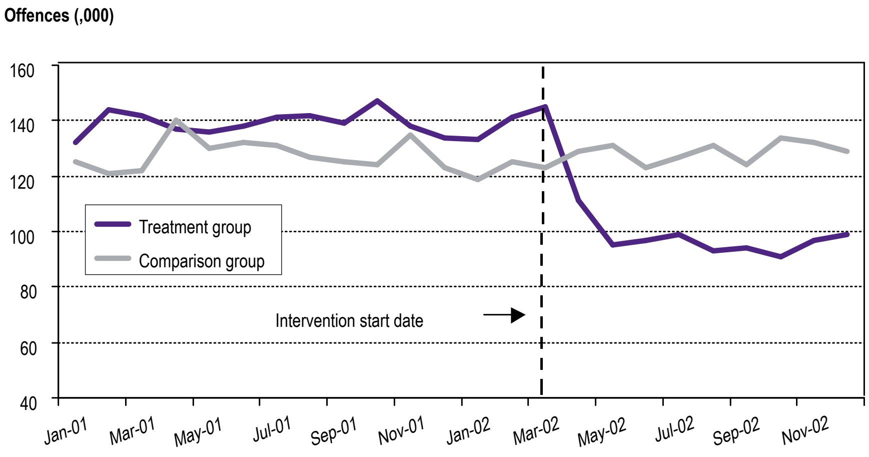
The best way to evaluate a program is to conduct a randomised controlled trial (RCT). In an RCT (see Box 1), individuals are randomly allocated to treatment and control groups. The treatment group receives the intervention whereas the control group does not. If both groups are large enough we can be sure that there are no systematic differences between them other than the fact that one group received treatment and the other did not. This means that any difference between them in terms of outcome can safely be attributed to treatment.

If an RCT is impracticable (as it often is), it is sometimes possible to conduct what is called a quasi-experimental study. There are several forms of quasi-experiment but the most powerful of them is known as the replicated interrupted time series design (Hall 2005). This design involves the use of both a baseline and a comparison group. An example of this design is provided in Figure 1.

In the replicated interrupted time series design (see Box 2), a series of observations of the outcome are taken from both treatment and control groups prior to the introduction of the program. The program is then introduced and a second set of observations taken from treatment and control groups. If the program is working, we expect the second set of observations to move in a more favourable direction for the treatment group than for the control group. Notice that by comparing outcomes for both treatment and control groups after the program we are able to control for any factors that might be influencing trends in both treatment and control groups.

It is sometimes impracticable (for ethical or practical reasons) to conduct an RCT or a quasi experiment. When this happens, the next best alternative is to try and match each member of the treatment

Figure 1: Hypothetical example of a replicated interrupted time series design



Box 1: Example of a randomised trial (Lind et al. 2002)

The NSW Drug Court took advantage of the fact that places on the Drug Court program were strictly limited. Each week the Drug Court would draw up a list of eligible offenders and then hold a random ballot to determine who among those eligible for placement on the program would actually be accepted on to it. Those who were placed on the program were included in the treatment group. Those who were deemed eligible for the program but who did not make it through the ballot were placed in the control group. The evaluation found that treatment group was slower than the control group to be reconvicted of a drug or theft offence.

Box 2: Example of a replicated interrupted time series design: Chikritzhs et al. (1997)

Chikritzhs et al. (1997) evaluated the impact of extended trading permits (ETPs) on assaults on licensed premises in Perth. They examined trends in assault on licensed premises before and after the introduction of ETPs and found the number of assaults on licensed premises went up after they obtained ETPs. This could have happened, however, simply because assaults went up on all licensed premises. To control for this possibility, Chikritzhs et all included a control group of licensed premises that did not obtain an ETP. No increase in assault was observed in these premises. This provided strong evidence that ETPs caused the increase in assaults and that it was not just the result of some pre-existing trend.

Box 3: Example of a matching study using statistical techniques: Weatherburn et al. (2008)

Weatherburn and Trimboli (2008) wanted to know whether offenders given a bond with supervision were less likely to re-offend than offenders given a bond without supervision. Offenders at higher risk of re-offending, however, are more likely to get a bond with supervision (selection bias). To get around this problem, they constructed a statistical model that allowed them to predict the likelihood that an offender would get a supervised bond. Each offender who actually received a supervised bond was then matched with another offender who did not get a supervised bond but who was just as likely (according to the model) to get one. The reconviction rates of these matched groups of offenders were then compared to see whether offenders who received a supervised bond were less likely to re-offend. No evidence emerged that offenders given supervised bonds were less likely to re-offend.

group with someone similar who does not receive the treatment and then compare the outcome(s) of interest for both. The matching can be done in a variety of ways but the most common method involves the use of special statistical techniques (see Box 3). This design is strong when we know what factors we have to match the treatment and control groups on but can cause problems when we don't, particularly if program managers or program gatekeepers select people for treatment according to who they think is most likely to 'succeed'. This problem (known as 'selection bias') is a major problem in criminal justice evaluation.

PRINCIPLE 3: THE BIGGER THE SAMPLE SIZE THE BETTER (WITHIN REASON)

Finding out whether a program 'works' is analogous to searching for a signal in a lot of noise. The 'signal' is the program effect. The noise is all the variation in the outcome that is due to factors other than the program. The larger the sample (i.e. the longer the baseline and follow up periods or the larger the treatment and control groups), the easier it is to separate the signal from the noise.

We noted earlier (see Principle 1) that program designers ought to state the size of the change in the outcome they expect or want their program to have. This is important for three reasons. Firstly, larger samples are necessary to detect weak signals (small program effects) than to detect strong signals (big program effects). Secondly, since the cost of many evaluations is closely related to the amount of data (size of sample) required, programs that are intended or expected to produce weak effects will often take more time and cost more money to evaluate than programs that are intended or expected to produce big effects. Thirdly, if we know what size signal (program effect) we are looking for, we can determine precisely what size sample we need to detect it. This saves time and money.

The 'within reason' clause is in our principle for two reasons. Firstly, sample sizes are subject to the law of diminishing returns. As the sample size gets larger and larger, the benefits of increasing the size of the sample still further get smaller and smaller. Past a certain point, increasing the sample size will add to the cost and duration of a study without adding appreciably to the power of the study to detect an effect. Secondly, if we are using a baseline, the further forward or backward in time we go to boost our sample size, the greater the risk that other unmeasured factors will create spurious results.

PRINCIPLE 4: PROGRAM FIDELITY IS JUST AS IMPORTANT AS PROGRAM DESIGN

The task of designing a program and getting its enabling legislation through Parliament is often such an exhausting task insufficient thought is given to program implementation. This is a great pity for, as the saying goes, the best laid plans of mice and men go oft awry (see Box 4). Many well-designed programs are either poorly implemented, not implemented the way they were planned or not implemented at all. As a general rule, the more complex the program and the more people involved in its implementation, the greater the risk of implementation failure.

The problem of poor program implementation has big implications for program evaluation. If the outcome evaluation is negative and no process evaluation has been conducted we don't know whether the program failed because it was poorly implemented or because it was a bad idea to begin with. As a result, all the effort involved in designing the program and getting its enabling legislation through Parliament has for all intents and purposes been wasted. Wasted too, are the efforts of those who worked diligently to put the program into place. This is why it is important to conduct process and outcome evaluations when implementing complex, expensive or risky programs.

Box 4: Operation 'Vendas' - A program with implementation problems (Jones et al. 2004)

Some years ago the NSW Police decided to evaluate a new program designed to reduce motor vehicle theft and burglary by making more effective use of fingerprint and DNA evidence to apprehend offenders. The program was implemented in three Local Area Commands (LACs), with all other LACs functioning as a control group. No benefits in terms of reduced crime were observed but the process evaluation revealed that there was very little change in the collection of DNA and fingerprint evidence by police in two of the 'treatment' LACs.

FIVE PRACTICAL TIPS

In light of the above, here are five tips for anyone wanting to commission an outcome evaluation.

TIP # 1: GET THE EVALUATOR IN EARLY

Do not wait until the program has been designed and key decisions have been made about program implementation (where, when and how). If the evaluator is not brought in early, it may be impossible to establish a baseline or organise recruitment of large enough samples. In this case evaluation will be impossible.

TIP # 2: THINK ABOUT HOW THE EVALUATOR MIGHT MEASURE YOUR PROGRAM OUTCOMES

What would change in the world if your program were successful? How could these changes be measured? If they can't be measured directly, is there something that could be used as an indicator that positive change has taken place? The evaluator will have ideas about this but if the program designer has no idea what a successful outcome would look like, the evaluator is likely to struggle too.

TIP # 3: FIND OUT WHETHER INFORMATION ON YOUR KEY OUTCOMES IS BEING ROUTINELY AND VALIDLY RECORDED

If you need a baseline and key outcomes are not being routinely monitored, it will be necessary to set up a system for measuring the outcome and establishing a baseline. This will take time and cost money. It may also mean that the program start date has to be delayed.

TIP # 4: THINK ABOUT HOW A CONTROL GROUP MIGHT BE CONSTRUCTED

Can you subject a group of people to the eligibility screening process for a program but not put them all on the program? Is there a surplus of people relative to places on the program that could be placed in a control group? Is it possible to collect the same information from a potential control group as from the treatment group? What information would you need to collect to create a group identical to the treatment group in all respects other than treatment.

TIP # 5: APPOINT A PROGRAM IMPLEMENTATION MANAGER

So the enabling legislation is through Parliament and everyone heaves a sigh of relief but whose task is it to ensure that the Government's plans are put into effect the way they were intended? It is surprising how often there's no simple answer to this question. Without a program manager, though, it is always difficult and sometimes impossible to put the administrative arrangements for evaluation into place. Do yourself a favour. Whenever a program is complex, expensive or politically risky, appoint a program implementation manager.

ACKNOWLEDGEMENTS

Special thanks to Rohan Lulham, Clare Ringland and Laura Rodwell for their very helpful and constructive comments on an earlier draft of this bulletin.

REFERENCES

Chikritzhs, T Stockwell, T & Masters, L 1997, *Evaluation of the public health and safety impact of extended trading permits for Perth hotels and nightclubs*, National Centre for Research into the Prevention of Drug Abuse, Curtin University, Perth.

Jones, C & Weatherburn, D 2004, *Evaluating Police Operations (1): A process and outcome evaluation of operation Vendas*, NSW Bureau of Crime Statistics and Research, Sydney.

Hall, R 2005, *Applied Social Research: A guide to the design and conduct of research in the 'real world'*, School of Social Science and Policy, University of New South Wales.

Lind, B, Weatherburn, D, Chen, S, Shanahan, M, Lancsar, E, Haas, M & De Abreu Lourenco, R 2002, *NSW Drug Court evaluation: cost-effectiveness*, NSW Bureau of Crime Statistics and Research, Sydney.

Weatherburn, D & Trimboli, L 2008, Community Supervision and Rehabilitation: Two studies of offenders on supervised bonds, *Crime and Justice Bulletin* 112, NSW Bureau of Crime Statistics and Research, Sydney.

NSW Bureau of Crime Statistics and Research - Level 8, St James Centre, 111 Elizabeth Street, Sydney 2000
bcsr@agd.nsw.gov.au • www.bocsar.nsw.gov.au • Ph: (02) 9231 9190 • Fax: (02) 9231 9187
ISSN 1030 - 1046 • ISBN 978-1-921626-06-7