Criminal Justice Evaluation Framework (CJEF): Conducting effective outcome evaluations

THE CRIMINAL JUSTICE EVALUATION FRAMEWORK (CJEF)

The Criminal Justice Evaluation Framework (CJEF) comprises a series of documents designed to introduce people who are inexperienced in evaluation to the various methodologies available so that they can determine the most appropriate approach for their particular evaluation. It is also designed to provide guidance to government departments and external agencies on the standard expected across the Queensland Government of evaluations conducted in a criminal justice context. While it is acknowledged that evaluating programs, policies, initiatives, interventions and operations that exist within real-world contexts requires flexibility, patience, and the capacity to communicate effectively across multiple disciplines and with persons inexperienced in research methodologies, high standards of research in all criminal justice evaluations are vital.

INTRODUCTION

Criminal justice agencies are increasingly being required to demonstrate the effectiveness of crime prevention, diversion and rehabilitation initiatives, and new service delivery models. Evidence of program effectiveness is critical to inform resource allocation decisions and to distribute limited funds across a highly competitive criminal justice sector. Yet few policy advisors and program coordinators have the time to develop significant expertise in research methods and design. This sometimes means that well-intentioned evaluations become methodologically flawed, making it difficult to meet Government requirements to provide evidence of a program's effectiveness.

Fortunately, policy advisors and program coordinators do not require a detailed knowledge of research methods and design in order to conduct meaningful evaluations and assist in the process of interpreting and critiquing program outcomes. Simply learning the basic principles of evaluation can help avoid costly mistakes and better demonstrate the outcomes of interventions. This document presents a brief introduction to the components of an effective outcome evaluation.

WHAT IS AN EVALUATION?

Evaluation is the systematic collection and analysis of information to make judgements about the effectiveness, efficiency, and appropriateness of a program or initiative. The principles of evaluation can be applied to many contexts (e.g. policies, programs, initiatives, interventions and operations). For ease of communication this document will broadly refer to these contexts as program evaluations. A program typically contains a series of actions or processes designed to produce some level of measurable change or outcome (e.g. an intervention to increase students' awareness of risk taking behaviour or police operations to increase arrests for alcohol related violence). Evaluation is a dynamic process that assists in the ongoing development and adaptation of programs to better suit the situational context in which they operate. Therefore, it is of benefit to policy advisors and program coordinators to incorporate an evaluation strategy in the early stages of program planning.

WHAT IS AN OUTCOME EVALUATION?

Outcome evaluations examine whether a particular policy, program, initiative, intervention, or operation (hereafter referred to as program) achieved what it intended to achieve. For example, an outcome evaluation might be conducted to determine whether the introduction of a need exchange program reduced the spread of blood-borne diseases among a prison population over time. Outcome evaluations rely on scientific methods to systematically measure both the positive and negative consequences of a program. This

type of evaluation rarely provides insight in to the reasons why a program was more or less effective in achieving its intended outcome/s. For that you will need to conduct a process evaluation.[1]

HOW TO CONDUCT AN OUTCOME EVALUATION?

An outcome evaluation is conducted in five steps:

1. Clearly identify the key research questions
2. Choose the most appropriate research design
3. Identify what data you will need and how it will be analysed
4. Adopt an ethical approach to evaluation
5. Clearly communicate your findings to others.

TIP #1: CLEARLY IDENTIFY THE KEY RESEARCH QUESTIONS

The success of an evaluation depends on the clear identification of the evaluation goals and key research questions. In other words, you need to know what you are trying to achieve before you try to achieve it. Typically, the goal of program evaluation is to determine whether a program has produced some level of measurable change. Identifying the goal of the program, therefore, is one way of identifying the goal of the evaluation. For example, if government introduces an alcohol awareness program in the hope of reducing alcohol related violence and improving community safety, the goal of a program evaluation is likely to be to determine how successful the program has been in reducing violence and improving safety.

Having identified the broad aims of the evaluation you can now begin developing clear research questions to achieve your goal. Good research questions are specific, directional, neither too broad nor too narrow, contain variables that are measurable, and can be answered in the time available to you. Table 1 provides some examples of good and bad research questions.

Table1. Components of a research question

Component	Good	Bad	Reason
Specific	Has the number of alcohol related assaults decreased since the introduction of the program? Has community attitude towards personal safety in entertainment precincts improved since the introduction of the program?	Has there been a decrease in alcohol related assaults and does the community feel safer as a result?	Including more than one variable in a research question heightens its complexity and increases the likelihood that components of the question will not be addressed.
Directional	Since the introduction of the program, has the number of alcohol related assaults reduced?	What has happened to the incidents of alcohol related violence since the introduction of the program?	The research question should state the direction of change you expect to see as a result of the program.
Neither too broad nor too narrow	Has the number of alcohol related assaults decreased in the two years following the introduction of the program?	Has there been a change in alcohol related violence? How many persons aged 21 years were charged with common assault in 2002?	Too broad. This question does not specify the type of violence, the direction of change expected. Too narrow. This question can be answered with a single statistic.
Variables that can be measured	On average, was there a decrease in number of alcoholic drinks participants consumed in a week after they had participated in the program?	After participating in the program, were participants more aware of the consequences of their actions?	It is not clear how we would measure an individual's awareness of the consequences of his/her actions. It is more appropriate to think about the type of behaviour a person would engage in if they were aware of the consequences of their actions. Behaviour can be measured.
Can be answered	In the three years following the implementation of the community	In the three months following the implementation of the community	It is important to consider whether it is reasonable to expect to see a change in

[1] For further information regarding this type of evaluation see *Criminal Justice Evaluation Framework (CJEF): Evaluating process and implementation.*

in the time available to you	education program, has the number of alcohol related assaults decreased?	education program, has the number of alcohol related assaults decreased?	outcomes in the given timeframe.

TIP #2: CHOOSE THE MOST APPROPRIATE RESEARCH DESIGN, COMPARISON GROUP AND SAMPLE

There exists a wide range of research designs from which to choose (e.g., experimental; quasi-experimental; naturalistic; ex post facto research; financial or economic analysis). Making the right choice will determine the success of an evaluation process. In particular, it will determine the extent to which your evaluation can be said to be internally and externally valid. Internally valid research is that which is able to show conclusively that an independent variable, in this case participation in a program, caused a particular outcome. Externally valid research is that for which the results are true across persons, circumstances, and time not included in the original experiment. Ideally an evaluation will contain both high internal and external validity. Realistically, however, the extent to which your research is internally and externally valid will be restricted by the nature of the populations from which you are able to draw your sample, the data you are able to access, and the timeframes and funding available to you.

The success of the evaluation process will also be determined by your capacity to adequately demonstrate an effect. Demonstrating an effect relies on the presence of an appropriate group with which to compare your participants. Only when you compare the outcomes for your program participants with the outcomes for persons who did not participate in your program, can you be confident that the changes you observe are due to your program and not some other influence (e.g., co-occurring initiatives). Selecting an appropriate comparison group will be restricted by the context in which your research is conducted. However, a comparison group should always be as similar to your test group as possible – particularly with respect to characteristics shown by the research literature to contribute to the outcomes of your program (e.g., age and gender distribution, and offence history). See Table 2 in order to determine the most appropriate comparison group in your situation.

Table 2. Methods of comparison ranked in order of highest to lowest control

Maryland Scale[2]	**Example of comparison method**	**Method**	**Advantages**	**Disadvantages**
Level 5: Random assignment and analysis of comparable units to intervention and control groups	Randomised Control Trial	Randomly assign persons from the population to participate in either the experimental (i.e. program) or control group Note. Random assignment is a specific technique which relies on random number tables to assign participants to groups – it is more than just drawing names from a hat	Reduces the risk of bias (e.g., only those persons who are likely to show the greatest improvement are selected for the experimental group) because all members of the population have an equal chance of being selected to be in the experiment or control groups	It is not always ethically or practically appropriate to use random assignment as a means of allocating persons to groups
Level 4: Comparison between multiple units with and without the intervention, controlling for other factors or using comparison units that evidence only minor differences	Matched Pairs	Identify persons to serve as a control on the basis of characteristics shown by the research literature to contribute to the outcomes of your program You may match at the level of individuals (for every individual in the test group who has a particular constellation of characteristics there is a matched individual in the control group) or at the group level (the distribution of key characteristics is the same across both experimental and control groups)	Improved sensitivity to change over other control techniques by ensuring that program participants and their comparison persons are similar	It is sometimes difficult to identify those variables on which it is most important to match participants The more variables on which individuals/groups need to be matched, the harder it is to build an adequate comparison Some variables are relatively rare/unique meaning that it can be difficult to find an appropriate match When matching on the basis of groups, the relationship between

[2] The Maryland Scale of Scientific Methods was developed by Sherman and colleagues (1998) as a means of classifying the methodological strength of studies evaluating crime prevention programs.

		Potential candidates for this method of control include: persons for whom you have data on the variables of interest that was collected prior to the implementation of the program; persons who have registered interest in the program and have been placed on a waiting list; persons who are ineligible to participate for reasons other than those likely to contribute to program outcomes		variables of interest may differ between groups (e.g., age & criminality) even though the distribution is equivalent
Level 3: A comparison between two or more comparable units of analysis, one with and one without the intervention	Spatial Comparison	Responses collected within one geographical location are compared with responses collected in a similar location	It is possible to show if, and how, participants' responses are different from responses given by persons who reside outside the test area	Without controlling for naturally occurring differences between residents in the comparison locations it is impossible to know whether the program alone or the identified differences are responsible any changes observed
Level 2: Temporal sequence between the intervention and the outcome clearly observed; or the presence of a comparison group that cannot be demonstrated to be comparable	Temporal sequence design[3] (Pre & Post analysis)	Responses collected prior to the introduction of the program (baseline data) are compared with responses collected once the program has been introduced Temporal comparisons may compare responses from the same participants over time or responses from different groups selected to be as similar as possible on the key variables (e.g., state wide offence rates in the two years prior to the change in legislation and the two years after the change in legislation)	It is possible to show if, and how, participants' response to specific measures/outcomes have changed over time	Without an appropriate control group (ie matched persons who did not participate in the program) it is difficult to know whether the change in participants' responses is due to the passing of time or the effect of the program
Level 1: Observed correlation between an intervention and outcomes at a single point in time	Post-intervention survey	Survey conducted with persons who participated in the program	It is possible to evaluate participants' unique experiences following involvement in a program	It is impossible to conclude the extent to which involvement in the program produced any level of measureable change on the outcome variables

The design you choose and the size of the intervention will also impact on the nature of your sample. If you are evaluating a program with only a small number of participants (e.g., workshop participants) then it may be appropriate for you to collect data from each participant. However, if you are evaluating a program with a large number of participants (e.g., persons arrested for public nuisance offences) you will need to draw a sample from the broader test population.

There are many ways in which you might select your sample (e.g., random, snowball or purposive). When selecting a sample there are two important things you should remember. The first is that your sample should be representative of the population from which it is drawn. The second is that your sample should be of an appropriate size to detect change, should it occur. If the evaluation sample is not proportionately representative of the population from which it is drawn, you will not be able generalise your results from the sample to the population.[4] Furthermore, you will not be able state with confidence that the effects you

[3] It may be important to note that temporal sequence design is different from time series analysis. The former is a research design while the latter is a method of statistical analysis. In particular, time series analysis uses statistical techniques, such as frequency-domain methods and time-domain methods, to interpret recurring patterns in data collected at regular intervals over time. That is to say, time series analysis comprises more than simply measuring differences between measures taken before and after a specific event.

[4] If your sample is not randomly selected so as to be representative of the population you risk selection bias. Selection bias occurs when some members of the population from which you are sampling have a greater likelihood of being included in your sample than

observed were due to the program and not some other factor unique to your particular sample (e.g., younger in age than the population). If your sample size is too small you may not detect changes that occur; if your sample size is too big, your analyses may be overly sensitive to change such that even small changes appear significant when in fact they are not.

Optimal sample size

The bigger your sample size, the better your chances of finding an effect (if it exists). The number of participants you recruit, however, will be determined by the resources you have available (time and money). One of the determinants of optimal sample size therefore is often the trade-off between the number of participants needed to show an effect and the resources available. Effect size refers to the amount of change you expect your program to produce in your outcome variable. For example, a school-based intervention (program) might be expected to decrease children's risk taking behaviour (outcome variable) by 10%. If you know the expected (or desired) effect size for your program (e.g., 10%) and rely on established thresholds for significance[5] ($p < .05$) and power[6] (0.8), you would need a sample size of 783 participants to confidently demonstrate this effect. The larger you expect your effect size to be, the fewer participants you will need to find the effect. In the above example, if we changed the expected effect size to 20%, the necessary sample size drops to 194; an expected effect size of 80%, reduces the sample size further to 9 participants. There are of course a number of ways in which sample size can be calculated (e.g., using known population size and desired confidence intervals). Table 3 provides examples of sample size calculations based on population size. Entering the search term 'sample size calculator' in your search engine will produce links to some automated calculators that can aid you in determining what is the most appropriate sample size in your situation.

Table 3. Sample sizes considered appropriate when drawn randomly from different populations (99% confidence level)

Population size	Sample size
200	171
500	352
1 000	543
2 000	745
5 000	960
10 000	1 061
20 000	1 121
50 000	1 160
100 000	1 173

Table adapted from Neuman (2006)

Note: The confidence level reflects how confident an evaluator can be that the responses obtained from the sample accurately represent the way in which the whole population would have responded if they had been tested. The above example shows the sample size required if an evaluator wants to be confident that 99% of the responses obtained are an accurate representation of the population's responses.

TIP #3: IDENTIFY WHAT DATA YOU WILL NEED AND HOW IT WILL BE ANALYSED

The data you collect will impact on whether you are able to answer your research questions. Data collected for the purpose of evaluations can generally be obtained from one, or more, of three sources: existing information; people; and observations. There are advantages and disadvantages associated with each data collection method which you will need to consider prior to making your selection. For example, while drawing from existing information may make data collection easier, these sources may not contain the information you require to answer your specific research questions. Furthermore, a significant proportion of the information may be missing. Conducting surveys with program facilitators may provide valuable insights into program processes, however, these methods are vulnerable to the influence of response bias (e.g., people respond favourably because they fear the consequences of responding negatively) and self selection bias (e.g., the experiences of sub-populations may not be captured if they all chose not to respond). The best

other members of the population. As a result any outcome you observe may be due to systematic differences between the sample and the broader population (e.g., program participants) and not the program.

[5] A measure of how likely it is that your results are due to chance and not the effect of your program.

[6] A measure of how probable it is that you correctly concluded your program had an effect.

approach is often to collect data from multiple sources as this allows for cross-validation of findings and builds a more thorough evaluation and makes findings much more valid. It is important to remember however that you should only collect data that will make a meaningful contribution in the quest to answer your research questions.

Once you have collected your data it will need to be analysed. The aim of data analysis is to synthesise information in order to make sense out of it. Therefore, it is important that you spend time considering which techniques are best suited to interpreting the data you have collected for the purpose of answering your research questions. Different analytical techniques are appropriate depending on whether you have collected qualitative or quantitative data. There are many statistical techniques you can use to analyse quantitative data. These techniques fall in one of two categories: descriptive (e.g., the number of participants who completed the program, or the rate of property offences) or inferential (e.g., Do reconviction rates of offenders who completed the program differ from those who failed to complete?) statistics. When analysing qualitative data, evaluators typically classify responses on the basis of re-occurring themes. Qualitative data is often used to provide deeper meaning to quantitative outcomes, while quantitative data is used to provide mathematical support to opinions expressed in qualitative data.

TIP #4: ADDOPT AN ETHICAL APPROACH TO EVALUATION

Adopting an ethical approach to evaluation is about more than doing what's right and avoiding what's wrong. According to the National Health and Medical Research Council (NHMRC), ethical conduct in research 'involves acting in the right spirit, out of an abiding respect and concern for one's fellow creatures.' In accordance with the NHMRC National Statement on Ethical Conduct in Research Involving Humans, any research or evaluation that *involves human participants* must be reviewed and approved by a Human Research Ethics Committee (HREC). These groups are established by institutions such as government departments, universities, hospitals, non-government organisations, and medical facilities to provide advice on ethical issues and to approve research projects.

Your role in gaining ethics approval depends on whether the evaluation is being conducted internally or externally. External evaluators will often gain approval themselves, although this will need to be stipulated in the evaluation tender documents and contracts. If the evaluation is internal to government, you will need to seek advice during the planning process of your evaluation about the specific procedures for gaining ethics approval within your agency. Adopting an ethical approach to evaluation ensures that participants are treated with the respect and protection they deserve and helps to build public trust and, subsequently, support for research outcomes. It can also help to ensure you comply with Information Privacy requirements.

TIP #5: CLEARLY COMMUNICATE YOUR FINDINGS TO OTHERS

Communicating your findings with conciseness and clarity will improve the chances of your recommendations being understood and implemented. Effective communication begins by identifying who is most likely to read your report (e.g., program stakeholders, those directly involved in the program or both) and their motivation for reading it. Knowing your audience and their needs will then help you to decide what information should be included in the report, the way in which it should be structured, and how to support the argument you are making (e.g., the program is effective, or the program needs to be changed). For example, it may be more appropriate for your findings to be released in a series of papers, each of which target the select needs of a unique audience than as a single report. Alternatively, your audience may respond better to workshop-style presentations or community meetings.

While the needs of your audience should play a key role in determining the way in which you communicate your findings and to some extent the level of detail you provide, there are some key components every evaluation report should include. These are:

- a description of the program being evaluated
- a statement containing the specific research questions addressed
- an explanation of the methods and measures used to collect the data
- the results found from the evaluation
- any limitations of the data, data collection, and evaluation
- a clear explanation of the answers provided to the research questions on the basis of the data collected
- any recommendations made on the basis of the results.

MORE QUESTIONS ABOUT EVALUATING?

If you are having trouble establishing a good evaluation framework or have any questions about evaluation please contact Criminal Justice Research, Department of the Premier and Cabinet (Ph: 32278436) or consult the references listed below.

REFERENCES

Australian Evaluation Society. (2006). *Guidelines for the ethical conduct of evaluations.*

National Health and Medical Research Council. (2007). *National statement on ethical conduct in human research.* Available at: http://www.nhmrc.gov.au/guidelines/ethics/human_research/index.htm

Neuman, W. L. (2006). *Social research methods: Qualitative and quantitative approaches* (6th edn). Boston, MA: Allyn and Bacon.

Sherman, L., Gottfredson, D., MacKenzie, D., Eck, J., Reuter, P., Bushway, S. (1998). Preventing crime: What works, what doesn't, what's promising, *National Institute of Justice Research in Brief*, U.S. Department of Justice, Washington.

Weatherburn, D. (2009). Policy and program evaluation: Recommendations for criminal justice policy analysts and advisors, *Crime and Justice Bulletin*, *133*.