College of Policing

NPCC National Police Chiefs' Council

Association of Police and Crime Commissioners

The Policing Evaluation Toolkit

Professor Stuart Kime, Director of Education, Evidence Based Education

Levin Wheller, Research and Analysis Standards Manager, College of Policing

Version 1.0

November 2018

This publication is available for download at: http://whatworks.college.police.uk/Support/Pages/Evaluation-Toolkit.aspx

Any enquiries regarding this publication please contact us at the College on 0800 4963322 or email contactus@college.pnn.police.uk

This document has been created with the intention of making the content accessible to the widest range of people regardless of disability or impairment. To enquire about having this document provided in an alternative format please contact us at the College on 0800 4963322 or email contactus@college.pnn.police.uk

Contents

Introduction

About this toolkit

This is a practical toolkit which brings together evaluation design and implementation strategies. It can be used by practitioners and researchers to ensure evaluations are designed in such a way that strong causal statements about effectiveness can be made.

We are keen to receive feedback from users of the toolkit to help us develop the content and ensure it is as useful as possible. If you have any feedback on the toolkit, please get in touch via email: research@college.pnn.police.uk

Evaluation and Evidence-Based Policing

Evaluation is a component of an Evidence-Based Policing (EBP) approach in which "police officers and staff create, review and use the best available evidence to inform and challenge policies, practices and decisions".[1] The College provides a range of resources to support people across policing and crime reduction to use evidence-based approaches.

What is evaluation?

Evaluation is a set of methods and tools which enable officers and staff to answer the questions 'Did it work?', 'To what extent did it work?', and 'Why and how did it work?' However, not all evaluations are the same: the quality of design and the execution matter.

Rigorous planning and execution are hallmarks of effective evaluation. In this toolkit, you will learn the methods of evaluation and how to apply them in your own context. Police forces, officers and staff can use this toolkit to better create, share and use evidence of what works, and test out new initiatives with a robust evaluative framework in place.

Key terminology: Intervention

In evaluation, an intervention is the thing that is evaluated. Any activity can be an intervention – often these are a specific tactic (or maybe a set of tactics), project or policy/policies introduced to solve a particular problem.

[1] The College of Policing definition of Evidence-Based Policing can be found here: http://whatworks.college.police.uk/About/Pages/What-is-EBP.aspx

How long does an evaluation last?

There is no single answer to this question. The length of an evaluation depends on the duration of the intervention being tested. Any intervention needs sufficient time to 'bed in' and for any effect it might have to be detected. However, evaluations should not go on so long that resources are wasted continuing with an intervention that may not have an impact.

What do we lose without effective evaluation?

Without robust evaluation, many advances in medical treatments, improvements in policing policies and effective education practices that have been witnessed in recent decades would not have been identified and used for public benefit. If we do not evaluate both the impact of a tactic, project or policy and the process undertaken to achieve it, there is no way of knowing if it was effective (see bottom right quadrant of Figure 1). In terms of maximising public benefit and reducing harm in a cost-effective manner, both impact and process evaluation are critical components of high-quality decision-making.

The 'Scared Straight' programme is one intervention that exemplifies the need for evaluation in the context of policing. 'Scared Straight' interventions aimed to deter children at risk of offending from offending in the future. However, evaluations found that, not only was 'Scared Straight' ineffective in its aim, it actually had a negative impact on the outcome of interest. It increased the chances of offending among the target population. Without evaluation, the true effect of a seemingly 'common sense' and intuitively 'right' intervention would not have been fully understood.

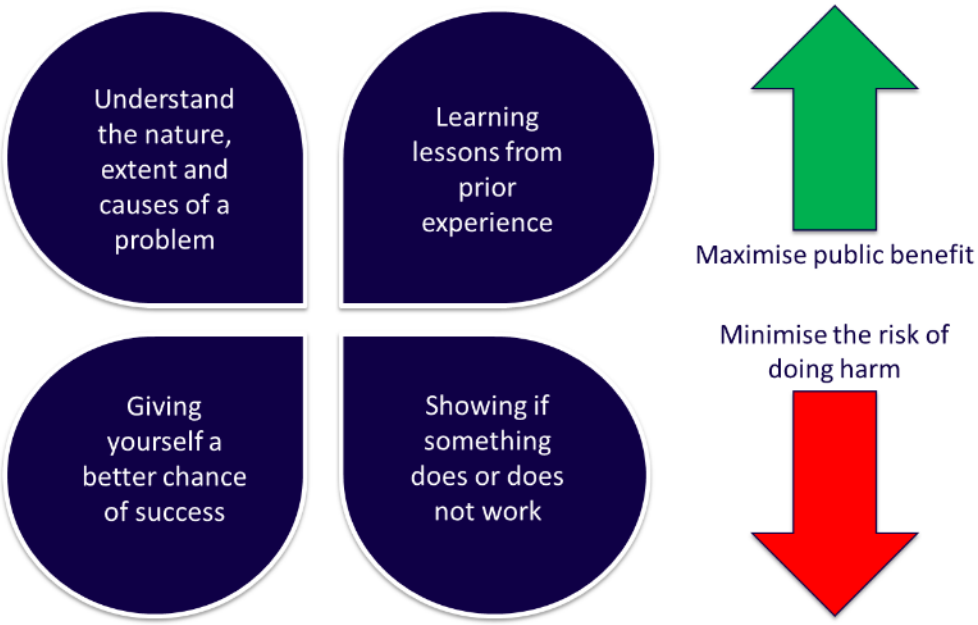

Figure 1: Benefits of an EBP approach

How can the Evaluation Toolkit help me?

- The toolkit provides a framework that can be used to help determine if an intervention is effective or not, and assess the strength of the impact.
- Effective evaluation can save time, resources and money. Using the Evaluation Toolkit to test tactics, projects and policies locally allows knowledge of effective and ineffective practices to be built up and used to inform decision-making.
- Effective evaluation guides future action and helps to ensure it is focused on where it will have the greatest public benefit. The investment of time and effort in designing and running an evaluation pays off repeatedly in better decisions about which activities to invest in.

How to use this toolkit

This toolkit should be used in **three stages throughout the life-cycle of an evaluation**. These stages are (1) preparation (2) implementation and (3) analysis and reporting.

Stage 1: Preparation (before the evaluation begins)

1. **Frame your evaluation question** – this is the question that your evaluation will set out to answer.
2. **Identify the measures you will use** – these are the indicators you will use to assess whether an intervention has been successful.
3. **Identify and recruit participants/locations** – create a sample of participants or select the locations you will use for the evaluation.
4. **Generate your comparison group or location** – to understand what would have happened to participants (or in the geographical area of interest) – if you did not implement the intervention. This is known as the counterfactual.

Stage 2: Implementation (during the evaluation)

1. **Conduct a baseline test** – to understand the starting point for participants or locations in relation to the outcome measure(s). Baseline tests can also help match intervention and comparison groups/locations (ideally these groups/locations will be starting from a similar place).

2. **Implement the intervention** – deliver the intervention as planned and record exactly what happens. You should ensure that your comparison group or location does not receive the intervention.
3. **Conduct the process evaluation** – to understand how the intervention is experienced by participants.
4. **Conduct a post-test** – to understand the impact of the intervention on the outcome measure. The post-test should be implemented at the same time with both the intervention and comparison groups/locations.

Stage 3: Analysis and Reporting (after the evaluation ends)

1. **Record and analyse your results** – the Evaluation Analysis spreadsheet can be used to calculate the effect of your intervention on the outcome(s) of interest.
2. **Report the results clearly and transparently** – finding a negative or no effect is as important and useful as finding a positive effect. Be sure to share your findings.

Key terminology: Baseline measures

Baseline measures tell you how things currently are, they give you a picture of the "as-is" situation. Baseline data are typically a measurement of the extent of the problem you are trying to solve taken before you start your intervention. To obtain a baseline you need to use data that can quantify the problem (e.g. recorded crime, case files or survey data). Data can then be used to measure changes in the problem over time, both during and after implementation of your intervention.

System change projects

Read this section if you are interested in evaluating a force-wide programme or a large-scale system change project.

The Evaluation Toolkit is designed to help officers and staff evaluate the impact of tactics, projects or policies in their local area. If, however, you are evaluating a system change project – such as the force-wide or national delivery of a new IT service – a slightly different approach is needed. Large-scale system changes often necessitate the delivery of a project to everyone at the same time, which means our ability to compare people's experiences of the 'old' and 'new' systems disappears. In such circumstances, the following guidance will help generate an understanding of the effect the project has:

1. **Pilot the proposed change with a small, representative group of users** – The piloting process could involve interviews/focus groups with users to understand more about their experience – this is often a useful sense check that can inform tweaks to the proposed plans and highlighting any room for improvement.
2. **Use baseline assessments/surveys (pre-tests) to understand the starting points of users prior to the change being delivered** – Gather data on those metrics the project seeks to address (e.g. time on task, response time or ease of use).
3. **Use follow-up assessments/surveys (post-tests) to gather data on the impact of the project on key metrics** – This will enable comparison of the baseline data with the outcome data. Although it is difficult to directly attribute any improvements to the intervention, the difference between pre- and post-test scores can help inform discussions and subsequent decisions.
4. **Use process evaluation** – User experience surveys, interviews and focus groups can help to understand not just what happened when the project was delivered, but how it was experienced by users.

The ways in which we intend changes to a system to be used or experienced are not always the same as what happens in the real world. Taking time to understand if people affected by your proposed changes share your understanding of the problem you are trying to solve is critical. Piloting your system change interventions and putting in place measures that help to understand more about how a project works when it moves from planning to implementation can be invaluable to achieving your aims.

Stage 1: Preparation

Preparation Stage 1.1: Frame your evaluation question

Asking a precise question is the first step to ensuring that an evaluation is successful; time put in to this part of the preparation stage is well spent and increases your chance of designing and implementing an effective evaluation. To frame the evaluation question, three choices need to be made:

1. **Intervention** – the new tactic, project or policy to be tested and the approach with which it will be compared.
2. **Outcome** – the outcome(s) you will use to measure the effectiveness of the intervention.
3. **Context** – where the intervention will be evaluated and the participants/locations that will be involved.

Examples of evaluation questions

- What is the impact of alley gating on the incidence of domestic burglary in student accommodation in Durham City?
- What is the impact of road side random breath tests at vehicle check points on the incidence of alcohol-related vehicle collisions among males aged 18-25 in the Trafford area of Greater Manchester?

Logic models

Logic models can help you to fine tune your evaluation question. A logic model is a step-by-step approach for clarifying responses to a specified problem which focuses on specific outcomes and how they will be achieved (the mechanisms involved). Developing a logic model in response to a well-defined problem will help to ensure your evaluation question is as specific and focused as possible. The College has developed a guide to logic models (see Figure 2). Logic models always begin with a clear understanding of the problem and can be used in conjunction with problem-solving approaches, such as the SARA (Scan, Analyse, Respond, Assess) model.

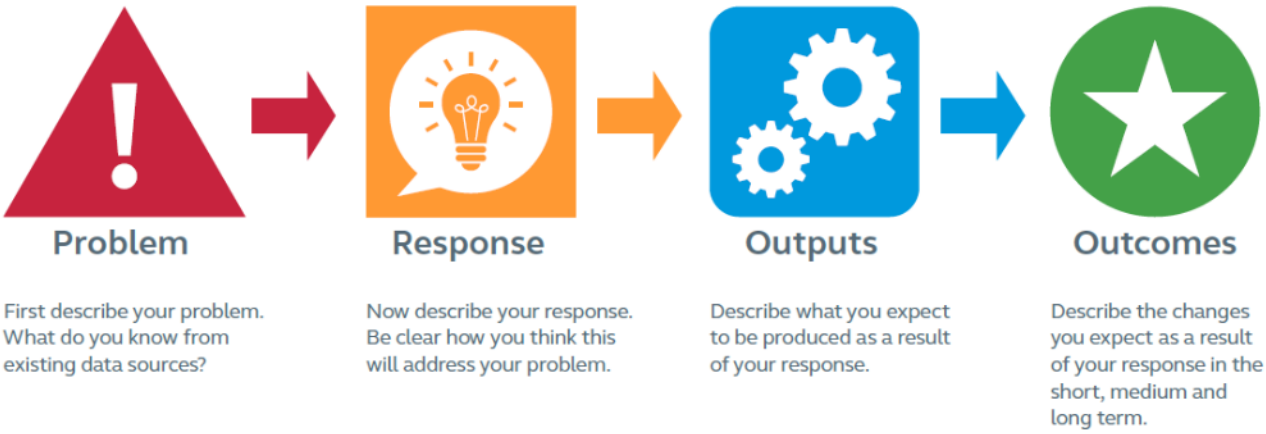

Figure 2: Overview of a logic model

Tips on framing evaluation questions

- Use a logic model to develop your evaluation question.
- Once you have drafted a question, refine it to make it as specific as possible. Vague questions are a major barrier to successful evaluations.
- Be clear about the **intervention** you want to test, whether this is a new approach or something that has been tested elsewhere before (e.g. from the Crime Reduction Toolkit).
- Be clear about the intended **outcome(s)** and how they will be measured.
- Be clear about the **context** in which you will implement your intervention. What people, settings or communities do you want your findings to apply to? Bear in mind approaches that work with one group may not work with another.
- Involve colleagues in the discussion – invite challenge to your thinking.

Preparation Stage 1.2: Identify the measures you will use

Once an evaluation question has been defined, the next step is to determine the measure(s) against which success will be judged. These measures will indicate whether the intervention has been successful.

Validity and reliability

The key objective in this step is to select a measure which is both **valid** (meaning it measures what it claims to measure and enables you to make the claims you want to make)

and **reliable** (meaning that it is both accurate and consistent over time and context). The box on the next page outlines seven tips for developing good outcome measures, which you can apply to any measure you are considering. If you cannot answer these questions satisfactorily then you should choose another measure.

Seven tips for developing good outcome measures:

1. Be sure you can clearly define what will be measured.
2. Be sure the proposed measurement instrument (e.g. survey, focus group) or procedure looks appropriate for measuring your outcomes.
3. Be sure you have considered what other factors (beyond what is supposed to be measured) might influence your outcomes? (For example, fear of reporting crime may influence any measure of incidents of crime.)
4. Be sure you have considered if other potential outcomes should be measured (e.g. how is public confidence in the police affected by a new policy, project or tactic). Ask yourself if there are any gaps in what you are measuring.
5. Be sure the measures aren't already collected elsewhere.
6. Be sure proposed outcome measures are dependable and trustworthy. If the measurement was repeated or collected by someone else, would you get the same results?
7. Check your thinking – ask colleagues to review and comment on your proposed measures.

Administrative data

It is important to consider what existing administrative data might be available to you, and whether this data can offer useful and appropriate outcome measures. Forces collect and hold a lot of data on crime, calls for service, intelligence reports, public satisfaction, and community priorities. It is always worth considering if you can use this sort of data in your outcomes, as this may mean you do not have to set up new data collection procedures.

Questions you should be able to answer about your evaluation

1. Precisely which baseline and outcome measure(s) will be used?
2. How and when will the measurements be taken?

Case study: Selecting measurements for a crime reduction intervention

You have been asked to evaluate a new approach to reducing theft of and from motor vehicles in your local area. The intervention involves two linked initiatives: target hardening (improving the security of vehicles and personal property) and raising awareness (leaflets and community engagement). Having selected your interventions you need to decide how you will measure its overall impact – there are several potential options:

Crime measurement – you will almost certainly want to use recorded crime as an outcome measure. Often you might only include the crimes directly linked to your intervention as your outcome measure, however it is possible your intervention might have an impact on other types of crime. Could encouraging people to make their cars more secure lead them to take similar steps to secure their homes and other property?

Public confidence – as well as measuring the impact on certain crimes, your intervention may also affect the local community's confidence in the police. Therefore, you may want to include public confidence as an outcome measure to see if your problem solving initiative has made an impact on this.

Victim experience – you may also want to measure victim satisfaction to understand how the introduction of your initiatives have impacted on victims of crime.

Unexpected impacts should be taken into account. For example, if public confidence increases in the local area because the initiatives are helping to reduce crime, you might find that calls to the police and reported crime goes up. Often as the police address and respond to issues which are important to a local community, trust in the police may increase, which can in turn lead the public to reporting more crime, as well as providing intelligence which could help you detect more crime.

Preparation Stage 1.3: Identifying and recruiting participants/ selecting locations

Recruiting participants

If you are recruiting individual officers, staff, victims or offenders to your intervention you need to decide and clearly state:

- Who are deemed to be 'eligible participants'? (e.g. 'young women in the West Midlands aged 15-18 who receive therapeutic foster care at the time of the intervention')
- What are the dates when recruitment of participants will take place?

Selecting locations

If you are implementing a location-based intervention, the intervention site can often be pre-determined based on a need to address a particular problem. It is, however, important to consider carefully:

- In which locations is it possible to implement the intervention?
- How will you find a suitable comparison site/location?
- What are the dates during which the intervention will take place?

Reducing bias

If possible, the identification and recruitment of individual participants should be done by someone who is not involved in the process of allocating which group each participant goes into (intervention or comparison); this is called 'blinding' and helps to reduce bias.[2] It is important to complete recruitment activities before the allocation of participants to groups, as this can also help to reduce bias. As such, good practice suggests that nobody should be added to the evaluation after the initial recruitment phase has ended.

[2] **Blinding** is where information about the assignment of participants to their experimental group (e.g. control or treatment) is concealed from the evaluator, the participants, and/or other people involved in the study until it is complete. A study is **biased** if its estimate of impact varies from the real impact. This variation can be linked to weaknesses in the implementation or design of the evaluation. See our research terms glossary for more information about these terms.

Involving stakeholders

When you are recruiting participants (individual officers and staff, for example) for your evaluation it can be useful to involve stakeholder groups during the early stages of the project. Early involvement can help to create a sense of 'ownership', which is useful in keeping people engaged over time (particularly in reducing drop-out rates during projects).

Making participation in a project easy, attractive, social and timely is a technique that has been used successfully in behavioural change initiatives and may offer useful tools for recruitment of participants.[3] Consider framing your recruitment activities in ways that are accessible to your target group; something as simple as encouraging participation through text message reminders can have a positive effect. Find out more by visiting the Behavioural Insights Team website.[4]

Preparation Stage 1.4: Generating a comparison group

After identifying and recruiting participants/selecting locations, the next step is establishing a comparison group or 'control' in order to understand the impact of the approach you are testing. The comparison group helps to overcome one of the key problems associated with making a causal link between an intervention and an outcome: the confusion of correlation with causation.

What happens when a comparison group is absent is illustrated in Figure 3 (below). In this example our evaluation question is: What is the impact of mobile devices on the time officers spend on visible patrol in our local force area? We have pre- and post-measures which show a change in time spent on visible patrol over time, so we can assume a link (a correlation) between the introduction of the devices and an increase in officer time on visible patrol. But how certain can we really be that the intervention itself has caused the change in our outcome measures?

Other factors might also play an important part in the observed change. One such example could be something as simple as the weather – if our pre-measure was taken in the depths of winter (when you might prefer to be inside, rather than on visible patrol) and our post-measure was taken in the middle of summer (when you might prefer to be outside), the change in time spent on patrol between the pre- and post-measures might be to do with

[3] http://www.behaviouralinsights.co.uk/publications/east-four-simple-ways-to-apply-behavioural-insights/

[4] http://www.behaviouralinsights.co.uk

changes in the temperature outside rather than the introduction of mobile devices. A comparison group or area allows us to compare changes in our intervention area to 'business as usual' – in other words it helps us see what would have happened without our intervention.

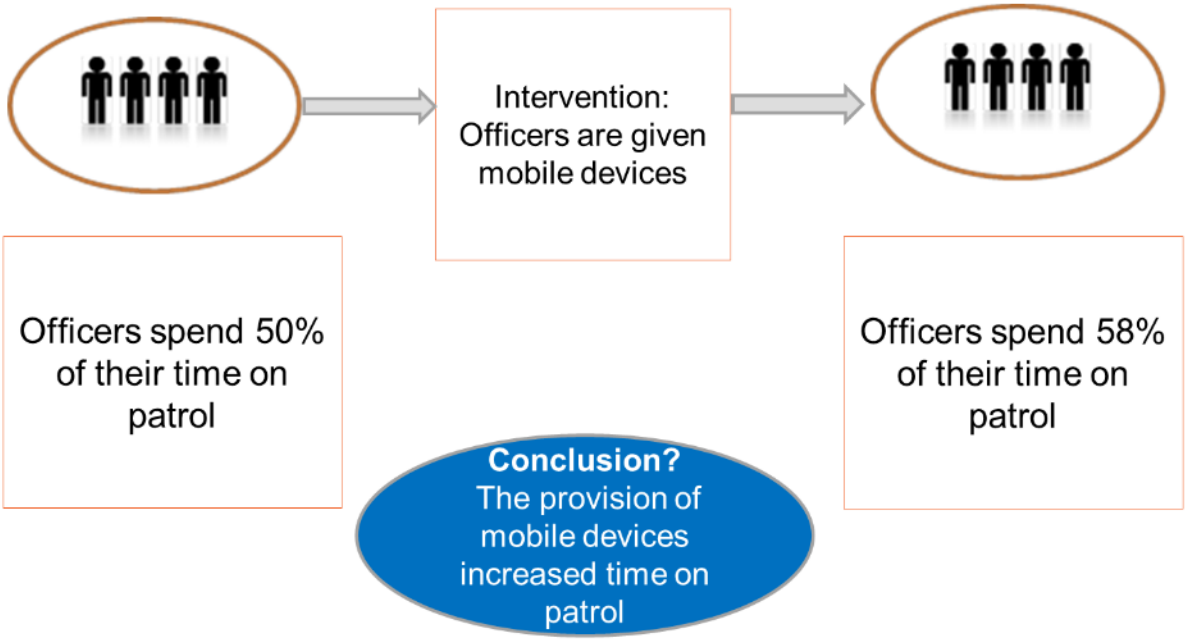

Figure 3: a common and ineffective way of identifying cause and effect

In order to get a more reliable understanding of the effect caused by a specific policing intervention, the inclusion of a comparison group (as seen in Figure 4, below) is required.

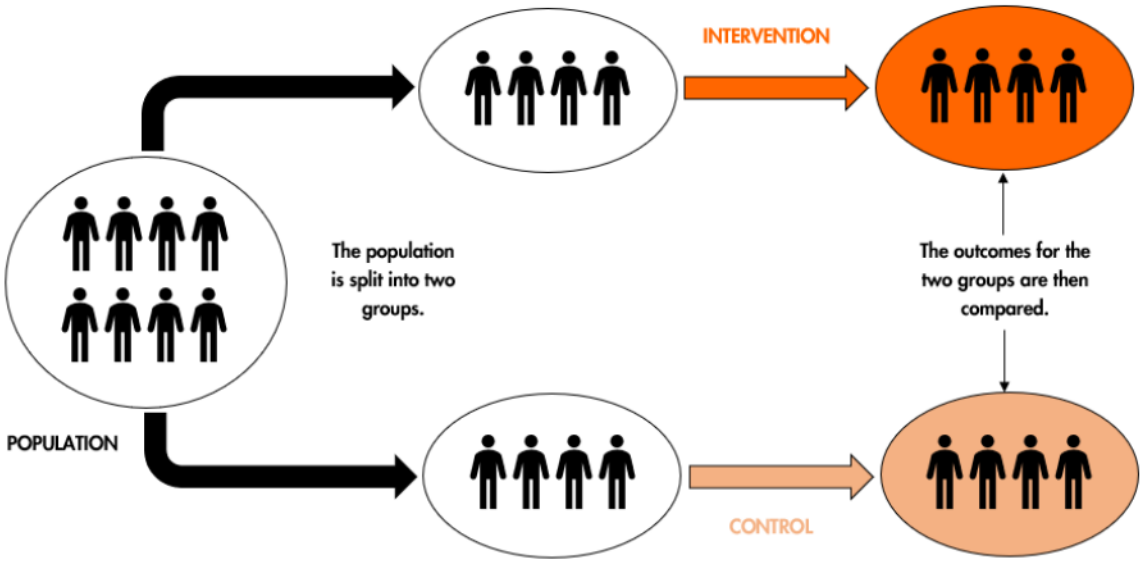

Figure 4: an evaluation design with the inclusion of a comparison ('control') group

The key factor in deciding to identify or generate a comparison group is whether or not you are in control of who will receive the intervention. If you do have control, it is recommended that random allocation (evaluation design A, below) should be used. If not, then it is recommended that matched control groups (evaluation design B, below) be is used. Other evaluation approaches – such as longitudinal studies with repeated measures over time – can also be used to make causal inferences, but these are not discussed further in this toolkit.

Evaluation design A: Random allocation

Random allocation is the best way to establish a comparison group – it provides the best chance of answering the question 'Did it work?' Deciding on the allocation of individual participants, groups of participants, or locations[5] to the treatment or control group at random guarantees that any initial differences between the groups result only from chance alone. Two short videos 'Why randomise?', and 'How can you randomise?' from J-PAL[6] and the World Bank respectively offer a useful overview of the value of randomisation.

Technical tip: simple randomisation using Excel

The RAND and RANDBETWEEN functions in Excel will produce a sequence of pseudo-random numbers that can be used in random allocation.

To randomise a dataset in Excel, assign each case a random number using the RAND function, and then sort the entire dataset by this column. Prior to the sorting process it is good practice to preserve the sequence of random numbers by pasting the random values into the dataset, since the RAND function will recalculate and generate a new set of random numbers every time the workbook is updated, overwriting the values you had previously. If possible, never use techniques based on birth dates, letters in the subject's name, every-Nth-case, etc., as the data may be stored in such a way to generate a biased sample.

[5] Randomising locations can be done in several ways, and the best way to do this will depend on the type of intervention you are implementing. The delivery of some tactical approaches can be allocated to random streets, crime hot spots, or wards, whereas partnership approaches may need to be delivered across whole ward or local authority areas.

[6] J-PAL is the Abdul Latif Jameel Poverty Action Lab, a global research centre working to reduce poverty by ensuring that policy is informed by scientific evidence.

If the groups are large enough and the random allocation has been done properly, it is extremely likely that the groups or locations will be equivalent with respect to every possible characteristic. Therefore, the only difference between the two groups is that one receives the intervention and the other does not. Practically, this means that at the end of the trial any observed difference(s) between the groups will be as a result of the intervention.

Without random allocation there are likely to be systematic differences between the groups (such as those who volunteer versus those who do not, or those who are enthusiastic about the intervention versus those who are not). When this happens, it is impossible to say whether it was these pre-existing differences or the intervention that made the difference.

There are several different ways of randomly allocating participants to the intervention and comparison groups. The choice of allocation procedure should be fit for the purpose and context of the evaluation. Five different approaches to randomisation are described in the box below. For advice on the most appropriate method for your evaluation, contact the College research team.

Five approaches to randomisation

1. **Business as usual:** The control group experiences things as they would have done 'normally', i.e. as they were before the intervention was proposed.
2. **Alternative treatment:** When the focus is on comparing two competing approaches, it is possible to allocate participants randomly to each group. If both approaches are plausible and acceptable, it might be fairer to allocate randomly than letting people choose which one they do. If choice of group is allowed, it is likely that one option may be more popular than the other and there are likely to be systematic differences in who chooses which version (for example one might be favoured by less vulnerable members of the public).
3. **Compensation:** Sometimes you might want to compensate participants in the control group by providing them with something different. When using this design, it is important that the compensation does not have an effect on the outcome you are measuring.
4. **Waiting-list design:** Here, everyone gets the intervention in the end, but random allocation (a lottery, essentially) decides who gets it first and who gets it later. The 'later' group acts as a control in the first phase. This design is particularly appropriate when constraints on personnel or resources mean that not everyone can receive the intervention at the same time: making the choice at random is not only fair but allows

the impact to be reliably estimated. This can also be useful design where there are ethical concerns about denying an intervention to some groups rather than others – eventually all people will receive the intervention.

5. **Border-line randomisation:** This can be used when an intervention is intended for specific groups, such as low-to-medium risk individuals. One group should definitely receive the intervention (those at high risk), another group definitely does not need it (those with no risk), but there might be a third group, such as low- to medium-risk individuals, for which you do not know whether giving them the intervention is the best use of resources. Participants in this borderline group can be allocated at random and their results compared.

Ideally, the participants/locations that are allocated (through whatever means available) to a comparison group should be as similar as possible to the intervention group in terms of observable, relevant background characteristics. For groups of individual participants, these include age, gender and geographical location. For locations, these include the crime profile and socio-demographic characteristics. Randomising people or locations to either the intervention or control group helps to account for these differences as well as those differences that we cannot observe (such as attitudes and feelings) – this is discussed in more detail in the next section on matched designs.

The comparison between a group of participants receiving the intervention and a similar but separate group continuing as normal allows us to estimate what would have happened without the intervention, a concept known as “the counterfactual”. We can never truly know exactly what would have happened, but a control group provides the best possible estimate.

If you are using a randomised design (evaluation design A), completing the following CONSORT Flow Diagram (Figure 5) is considered good practice, since it enables you to quickly check how balanced the intervention and comparison groups are. While not all of the boxes in the Flow Diagram can be filled in at this stage, it is a useful tool to help identify key elements that will need recording along the way. An example of a completed CONSORT diagram for a policing study can be found in the GMP Procedural Justice Training Experiment (p20).

Recap: 8 steps to conducting an evaluation using randomisation:

1. Register your research on the College of Policing Research Map.
2. Identify and recruit participants/locations (command areas, officers, offender/victim groups).
3. Collect baseline measurements.
4. Randomise participants to two or more groups (intervention and comparison).
5. Implement the intervention.
6. Collect post-intervention measurements.
7. Analyse post-intervention measurement data.
8. Report and publish trial results (ideally according to CONSORT criteria, Figure 5).

CONSORT
TRANSPARENT REPORTING of TRIALS

CONSORT 2010 Flow Diagram

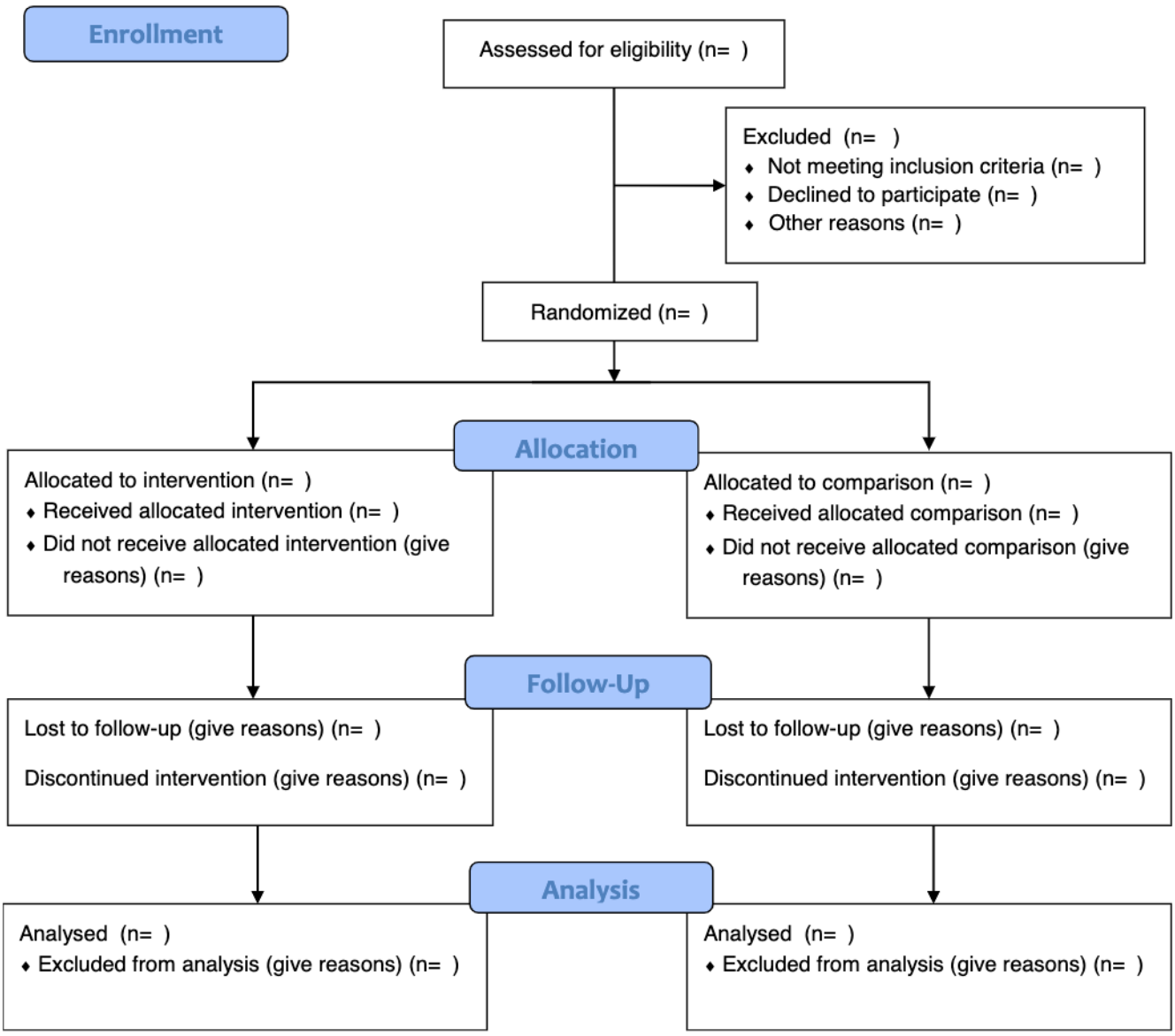

Figure 5: CONSORT 2010 Flow Diagram

Evaluation design B: Matched comparison groups and locations

Random allocation is the best approach to establishing a comparison group, but where this is not possible, a good alternative is to use a matched comparison group or area.

Matching groups of individuals

In this approach, the matched comparison group is made up of people who share key characteristics (e.g. age, gender, location) with those in the intervention group. It is essentially a data-based approach in which intervention participants are 'data matched' with other people who are broadly similar in as many important ways as possible.

The matched comparison group might be created in one of two ways:

1. from other officers or members of the public not receiving the intervention, or
2. from previous groups of individuals for which there is historic data using the same or similar measures planned for this evaluation (e.g. incidence of alcohol-related vehicle accidents).

The main limitation of this matching approach is that there are likely to be differences between the two groups that cannot be controlled for, such as socio-economic background, health and education.

Matching locations

If you are testing a geographically focused intervention it may be difficult to randomly allocate your intervention. Instead, you might undertake a simple comparison between your intervention area and another similar area in your force (or indeed another force). The locations you plan to use for treatment and control should be matched on some key characteristics (e.g. current crime profile, demographic characteristics and socio-economic characteristics). Poor matching may result in your evaluation highlighting only the pre-existing differences between the two locations, rather than the impact of the intervention.

Well-matched areas allow you to control for known differences using administrative area-level data (such as the average age of offenders, or the time of day/year at which incidents occur) and will give you greater certainty that any differences you observe though your evaluation are a result of your intervention. The main limitation of matching locations in this way is that there will still be differences between areas that even the best matching cannot account for. For example, there may be differences in the culture or attitudes of the officers and staff working in the two areas, which might influence how an intervention is implemented, and which in turn might influence evaluation outcomes.

When to use a matched design

You should use a matched design if:

- You are unable to use random allocation.
- You are able to establish a suitable local control group.
- You have access to good data on the initial characteristics of both groups.
- You have access to data on the same baseline and post-intervention measures for both groups.
- You are prepared to accept that the findings will be less reliable than in a randomised study, but can provide indicative evidence and foundations for further evaluation.

Ultimately, randomisation offers the best approach to understanding a causal link between an intervention and an outcome. If you want to find out more, there is a good explanation of why randomisation is important in Test, Learn Adapt: Developing Public Policy with Randomised Controlled Trials published by the Cabinet Office. Where randomisation is not possible, pre-post studies that utilise a well-matched comparison groups/areas to act as a counterfactual are a good alternative.

Questions to be answered before moving to Stage 2 (Implementation)

1. What is the background to evaluation (explain the rationale)?
2. What are the specific objectives of the evaluation?
3. What are the eligibility criteria for participants in this evaluation?
4. What are the settings and locations where the data are to be collected?
5. What is your evaluation question?

Once you have answered these questions, you are ready to move to Stage 2: Implementation.

Stage 2: Implementation

This toolkit now considers the design and implementation of effective evaluations.

Think 'EMMIE'

Viewing your planned evaluation in the context of the EMMIE approach[7] (see Figure 6) is useful in understanding the evaluation's place in the broader context of Evidence-Based Policing (EBP). Many evaluations often focus mainly on the effect of an intervention, but the EMMIE model helps ensure other important questions, such as 'How did it work?' (the mechanisms that helped achieve change) and 'How was it delivered?' (the implementation of the intervention), are properly considered.

EMMIE	What does it tell me?
Effect	What was the impact of my activity? Did it 'work'?
Mechanism	How does/did it work?
Moderator	Is the impact different in some areas/for some groups rather than others?
Implementation	Was my activity implemented in the way it was intended?
Economic cost	How much did my activity cost and what was the benefit?

Figure 6: The EMMIE model

This toolkit will lead you through a set of steps to enable you to answer all of the questions in the right hand column of Figure 6. Following these steps will help you develop a robust understanding, not just of what happened as a result of your intervention, but also of how the effect(s) happened (for instance, how participants experienced them). This part of the

[7] Bowers, K., Johnson, S., & Tilley, N. (2016). Introducing Emmie: An evidence rating scale to encourage mixed-method crime prevention synthesis reviews. Academy of Criminal Justice Sciences, Denver, USA.

evaluation – known as 'process evaluation' – is essential to EBP approaches. Without it, we would only ever have a partial picture of events.

Process evaluation should include the economic costs and benefits of your project, as outlined in the EMMIE model. You can get help calculating economic costs by using the College cost benefit tool.

Implementation Stage 2.1: Conduct a baseline test

Baseline measurements help to establish where participants (or locations) in the evaluation start from (in terms of characteristics relevant to the trial) and enable a good comparison group to be generated. You will have identified your measures in Preparation Stage 1.2.

How baseline data are useful

It is important that the intervention and comparison groups start from a similar point, therefore, you should conduct your pre-test:

- Before you start implementing the intervention you are evaluating
- At the same time for both the comparison and treatment groups
- At a time when as many of the participants as possible will be available (to ensure you have the largest possible sample for the analysis); and
- Before you randomly allocate participants to groups (taking baseline measures before the allocation process helps to reduce bias).

Ideally baseline measures (used for your pre-tests) will be the same as those used to measure the impact of your intervention (in your post-test). However, in some circumstances this is not possible (for example, where a new case recording system is introduced while you are planning your intervention) and you will need to use measures that are as similar as is practical.

It is generally best to collect individual participant baseline data, but aggregated group data are often good enough (especially where your outcomes will be analysed as averages across the group). Whatever constraints you work within, the most important thing is to be aware of the limitations of the data you collect and report these transparently when sharing your results.

Implementation Stage 2.2: Implement the intervention

Prior to implementing the intervention, it is critical to document exactly what is intended to happen (e.g. how long the intervention will last; how often it will take place; the training or preparation needed by those delivering it). Documenting your planned approach will help ensure that if the intervention is successful you will know exactly what you did to make it work. Your implementation plans are hugely valuable and can be shared with officers and staff in other forces who may want to replicate your approach.

Even the best laid plans can go awry, and often interventions may not be delivered exactly as intended. Those delivering an intervention may change it, select from it, improve it or just fail to do it properly (e.g. the plan might have been to deliver an intervention daily, but in reality it may have been delivered only once a week). In cases where your intervention does not appear to be effective, a process evaluation will ensure you can check whether this was because it was not implemented or experienced as planned. In cases where an intervention is effective, you may have learned something new (e.g. weekly delivery of the intervention might be just as effective as daily delivery).

Implementation Stage 2.3: Process evaluation

Alongside any impact evaluation it is important to undertake a process evaluation. This will help you to understand how your intervention was delivered on the ground, any challenges that were faced, and any improvements that might be made. Process evaluation can answer questions about your intervention including:

- Was it delivered as intended?
- What are participants' perceptions of the approach?
- What has worked well and what has not worked so well?

Information from the process evaluation will enable you to understand how the intervention might be improved and whether it is practical to roll it out more widely. There are various kinds of qualitative and quantitative data you could collect in a process evaluation. For example:

- Delivery records can help you understand how many sessions were actually delivered and to whom.
- Observations can help you understand how an intervention is being delivered in practice.
- Interviews with, or surveys of, participants and officers involved in delivery can help you understand their perceptions of an intervention and of any implementation issues.

Process evaluations can also act as a useful 'health check' for your intervention, to make sure things are progressing as expected. The case study example below illustrates how process evaluation can help identify – and rectify – issues with how your intervention is being implemented.

Process evaluation case study

A Chief Inspector found – through detailed analysis – that many of the burglary offences in his area were at repeat or near repeat locations and followed the same method: access from the rear using alleyways.

Whilst working on the long-term goal of getting alley gates fitted to prevent easy access to the alleyways, the Chief Inspector instituted a plan every time there was a break-in. The targeted house would get a 'Gold' service, (window locks, window alarms, property marking etc.) those adjacent a 'Silver' service (a reduced version of the 'Gold' package) and those other houses in the vicinity received a 'Bronze' service (leaflets through the doors to reflect what the evidence base said about the likely pattern of offending). The plan was to assess the impact of target hardening in preventing repeat/near repeat burglaries.

Enter one of the Chief Inspector's very keen and well-meaning Sergeants. At the end of week one he very proudly announced that he had used up all of the (expensive) locks and alarms by giving everyone on the street the 'Gold' service following a break-in. Clearly, the Chief Inspector had not explained things properly, the Sergeant was trying to help his community by doing what he thought was best.

A planned process evaluation would ensure that this action would have been picked up and quickly stopped, ensuring the intervention could from thereon be implemented as intended. Being aware of these sorts of issues in a project is also hugely important in interpreting results from an evaluation of the strategy.

In some instances it will be important to use ongoing process evaluation to check (and if necessary change) the way the project is being implemented.

You should carefully select the most relevant data or method to keep the process evaluation lean. You should also ensure, particularly when collecting qualitative data, that the person doing so is as independent as possible; anyone who is connected to the intervention and has put effort into making it work may find it hard to report neutrally on it. Finally, it is important that the methods you use to undertake a process evaluation do not in themselves change the intervention as a result of trying to record it. For example, observing every

aspect of the intervention could change the way it is delivered – this phenomenon is commonly known as the 'Hawthorne effect'.[8] Observations should be unobtrusive, done on a sample only, and balanced across the treatment and control groups.

Remember, process evaluation data are complementary to impact data; they are not a substitute for a good impact evaluation. Combining both impact and process evaluations will give a much more rounded understanding of an intervention and its effects.

Implementation Stage 2.4: Conduct a post-test

A post-test is used to understand the impact of your intervention. It is important to think about the timing of your post-test; you should think carefully about the length of time needed for the intervention to take effect and your post-test should be conducted after this point.

There is merit in conducting one post-test immediately after the intervention is completed and another after a further period of time. This second measurement will give you an indication of the longer-term effectiveness of your intervention and whether any impact is sustained.

Conduct your post-test:

- At the same time for both the comparison and intervention group; and
- At a time when as many of the participants as possible are available (to increase the likelihood of a large sample for the analysis).

If the post-test involves some human judgement, it is important to consider the extent to which the measurement of the outcomes could be influenced by the expectations or hopes of those taking the measurements and think seriously about how to 'blind' them to the identity of the participant (if possible) and which group (intervention or comparison) a participant is in. If the project team member responsible for measurement knows whether participants are in the intervention or control group, this may introduce bias to the evaluation and degrade its overall quality; this bias is subconscious and inevitable no matter how honest we think we are being.

[8] The Hawthorne effect (sometimes called "observer effects") is a phenomenon where participants change their behaviour due to the knowledge that they are being studied. For example, officers might follow a procedure or policy more closely if an evaluator is observing them on patrol. The presence of Hawthorne effects can lead to biased estimation of effect size. One way of avoiding the Hawthorne effect is to have a control group in your study.

There are two main ways to address this kind of bias:

1. Only use objective measurements: objectively-scored measures (such as durations of time and responses to multiple choice questions) do not require judgement, so are less likely to be prone to any adverse effects of human expectation.
2. Use ‘blind’ testing: having someone not involved directly with the evaluation thus far administer the post-test can help to reduce bias by virtue to them not knowing participants and the context. This may seem like a lot of trouble to go to and in some cases may not be possible. However, there is a lot of evidence that un-blinded judgement-based assessments are biased, often substantially so.

Some of these issues may be less likely to occur if your outcomes are routinely collected through administrative data or as part of ‘business as usual’ activity.

Stage 3: Analysis and Reporting

The final stage in the Evaluation Toolkit is analysis and reporting. The most appropriate analytical approach will depend on the type of data you have collected. For example, if you have numerical (quantitative) data, you will be testing for statistical differences. Remember, evaluations must compare data from at least two points in time to tell us anything meaningful about the impact of an activity.

Analysis and Reporting Stage 3.1: Analyse your data

Once you have completed your intervention and testing, you should analyse your data. One approach is to put all of your data into the Evaluation Analysis spreadsheet, which will then calculate an 'effect size' for your intervention.

Effect sizes

Effect sizes are quantitative measures of the size and consistency of the impact on an outcome. They are calculated by taking the average difference between two scores and dividing it by the variation in that difference.

$$\text{Effect size} = \frac{\text{Difference between average post-test scores}}{\text{The variation in that difference as a standard deviation}}$$

Interpretation

The Evaluation Analysis spreadsheet will help you interpret the impact data from your evaluation. The effect size data should be viewed in tandem with the process evaluation data (collected through delivery records etc.). The ultimate objective is to understand two things:

1. The effect the intervention had on the outcome.
2. The way in which the intervention had the effect.

Potential difficulties with interpretation

If you were able to randomly allocate participants/locations to groups and implement your evaluation exactly as planned, the only difference between the intervention and control groups should exist as a result of the intervention. Unfortunately, this is not always the case and there may be a number of reasons why differences have occurred.

When interpreting your data, you will need to consider the other factors that may have brought about the change (or lack of change) that you are seeing, including:

- **The intervention under consideration:** the effect on the outcome may be a direct result of the intervention you are evaluating.
- **Systematic differences between the groups:** if you were unable to use random allocation, there might be systematic differences between your groups that have brought about the effect. For example, one group of participants may have had the intervention implemented with them by a far more enthusiastic, more effective officer and reside in a place where they are implementing additional interventions which might affect your results.
- **Problems with your evaluation methods:** there are a number of factors regarding your evaluation that might affect your results. You should think about all the steps in this toolkit, and in particular whether there were any differences in the timing or delivery of your pre- and post-tests that might affect the results. For example, the intervention group may have been post-tested at a time when less participants were available than in the control group.

Clearly, many of the potential problems with interpretation can be reduced through sensible design and implementation decisions. Invariably, a successful evaluation is well-planned and faithfully-executed.

Analysis and Reporting Stage 3.2: Reporting your results

By now you should be able to tell if your intervention has had the expected impact. Did your new approach have the desired effect? Did it make a difference? What did you learn about implementation from your process evaluation?

What should you report?

It is important to report the results clearly so that others can understand what you did and what you found. No matter what the findings of your evaluation are, they should be reported transparently. Finding and reporting that an intervention has a negative effect (that it is

worse than the 'business as usual') is just as important as sharing data from a successful evaluation. There are a variety of ways to report your findings, and how you choose to do so will depend on who you intend to share your findings with. As a guide, Figure 7 outlines the broad overall structure of a typical research report.

	Section	Might include…
1	Title	Subtitle; names of report authors and affiliations
2	Executive summary (one page)	One page summary of the evaluation; background; main findings; discussion; conclusions
3	Contents	Page numbers for different sections of the report and any figures/tables/graphs
4	Introduction	Background to the evaluation; details of the intervention; details of the research aims; references to any relevant existing literature; your hypotheses
5	Methods	Research design; details about the selection of participants; details about the implementation of the intervention; details about how you have analysed your data
6	Results	Findings from the impact and process evaluations
7	Discussion	Discussion and interpretation of the findings; description of the practical implications of your findings
8	Conclusions/ Recommendations	Key conclusions and any recommendations arising from the evaluation
9	References	A list of any literature referred to in the report
10	Appendices	Space to include any further information that might be of interest, for example materials relating to the evaluation or your research/measurement instruments

Figure 7: Structure of a typical research report

Adapt your practice

When you have measured your outcome you should adapt your programme, policy or tactic response to reflect what you found. If you find that your intervention is not having any effect, then you should stop using the resources in this way and use the findings to adapt your understanding of the problem or change how you respond. If you find that your intervention is having the desired effect, then you might want to suggest trying it in other areas which are experiencing similar problems.

Share your findings

Throughout your evaluation, there is often considerable value in sharing your problems, plans, and findings. By sharing your questions and findings you might identify others trying to tackle similar problems or you can stop somebody using their resources on something you have found doesn't work.

Many forces now have local or regional partnerships with Universities to help support and promote research, examples include the Centre for Police Research and Learning, EMPAC, and the N8 Policing Research Partnership. When you start a project you can share your plans on the College of Policing Research Map or POLKA (the Police Online Knowledge Area) and you can get support through a College Research Surgery.

Sharing findings at the end of your project is important to help build the evidence base for the wider profession and ensure colleagues across the service can learn from your work. Organisations like the Society for Evidence-Based Policing can also provide advice and support, as well as the opportunity to share your work with colleagues across the service.