9

Evaluation models and methods
INTRODUCTION

Prior to developing an evaluation strategy and deciding on the best techniques to deploy in collecting information, it helps to clarify the purpose of the evaluation process. Data collection techniques are best considered as builder’s tools - deciding which tool to use will depend greatly on the job that has to be done. The choice of observation or focus groups or questionnaires will be guided by the objectives of the relevant evaluation exercise. Measuring the effectiveness of capacity-building programmes requires a process of evaluation that looks at what can be measured and how it can be measured. Since evaluation forms an integral part of the process by which capacity-building is designed, developed and delivered these measurements cannot be an end in themselves.

Figure 9.1: Step Six – selecting the most appropriate evaluation method

As observed from the outset, measuring the effectiveness of capacity-building must flow from a fully systematic and integrated approach to the development and the implementation of police training and development programmes. It follows that measuring the effectiveness of capacity-building need not take place at only one
stage, such as completion of a training and development programme. It can start at a number of stages by using data derived from the policing plan priorities and objectives, performance gap identification, job profiles, capacity-building design and existing capacity-building programmes. Each of these can represent a starting point for evaluation and generate a wealth of data to guide and structure the measurement process.

The key points of the chapter can be summarised as follows:

- Identifying the pivotal role of evaluation in a reform process
- Describing two of the most commonly used evaluation models
- Identifying some of the critical issues in evaluation design
- Identifying the three main methods of data collection and
- Describing how the principles of validity, reliability and relevance apply to evaluation data.

PIVOTAL ROLE OF EVALUATION

The traditional approach views evaluation as a process that only begins when police capacity-building has occurred and overlooks pre-existing data that can better ensure that evaluation is measuring what the organization demands. The Policing Plan should yield critical information that will assist in guiding the evaluation process. It should be scrutinised in search of answers to the following questions:

<table>
<thead>
<tr>
<th>Chart 9.1: Information derived from a Policing Plan</th>
</tr>
</thead>
<tbody>
<tr>
<td>▪ What is the police organization’s strategy under the Policing Plan and how does that strategy assist it in meeting its objectives?</td>
</tr>
<tr>
<td>▪ What is the position of capacity-building in the overall structure and function of the police organization?</td>
</tr>
<tr>
<td>▪ How does capacity-building support the organization’s strategy and objectives under the Policing Plan?</td>
</tr>
<tr>
<td>▪ What data does the organization need about the effectiveness of its capacity-building programmes?</td>
</tr>
<tr>
<td>▪ How will this data be applied?</td>
</tr>
</tbody>
</table>

The evaluation process demands clarity of purpose and the answers to these questions will yield a guiding framework for the measurement process. Where an evaluation lacks clear purpose it tends to wander aimlessly and results in findings that are devoid of organizational relevance.

Of course police organizations run at different speeds and levels of sophistication in terms of policing plans. In some there is no explicitly stated plan, in others the philosophy of working under the guidance of a plan is a new experience, and in others the organizational strategy may be subject to frequent change. Capacity-building specialists must be aware of these variations and be prepared to take a flexible approach in linking the evaluation process to the organization’s strategy. Specialists may otherwise make false assumptions about the nature of the strategy
at an operational level and thereby embark on an evaluation process that inevitably misses the mark.

Evaluation is vital in the process of ensuring the continuous improvement of the quality of the content and delivery of capacity-building programmes. It involves generating data through a process of inquiry and then, on the basis of this, making judgements about the strengths and weaknesses and the overall effectiveness of the course, and making decisions about how to improve it further. Specifically, evaluation will be used to address the appropriateness of the learning objectives and the length of the course, as well as assessing the effectiveness of the overall programme in terms of the performance needs it was designed to address. In addition to the use of evaluation as a quality assurance tool, it can also be used in response to a specific concern. In an organization that is faced with financial cutbacks or a rigorous budget-setting process, evaluation might be utilised to measure the extent to which an expensive capacity-building programme represents good value for money.

Figure 9.2: The role of evaluation in the police reform process

Equally, an evaluation can be used as an investigative tool in response to complaints about a capacity-building programme, or a device that will probe the effectiveness of a programme that is alleged to have failed to meet its intended objectives. Note that these various purposes are interlinked and closely related: whilst every evaluation will, in broad terms, focus on the effectiveness of a capacity-building programme the accent or emphasis can shift as needs require.

**Evaluation as part of the reform process**

Even though evaluation is a potent and necessary tool in police capacity-building programmes, it is often omitted in both internal programmes and donor-funded external programmes. This is an extraordinary failure, given that a considerable investment is allowed to seemingly avoid the quality and financial criteria that would be applied in the private sector. The net result tends to be a lack of rigour in future planning and a loss of credibility when poor programmes are repeated in spite of informal complaints from the beneficiaries.

Where evaluation is used, there is a second pitfall in the latent danger of it becoming a meaningless chore: familiarity can and does breed contempt when police officers are required to complete evaluation forms. The evaluation exercise can become routine for all the staff involved unless it is properly managed and subjected to constant review. An important factor in averting this danger is the need to stress the impact of the evaluation: i.e. its ability to make a difference. There is no point in conducting an evaluation for its own sake: it must be seen to result in data that will be used to make a difference to the quality of capacity-building and, in turn, the police organization’s performance. Herein lies the first and most important principle of evaluation – it must in every aspect be consciously linked to the aims of police reform.

There must be a continuous loop that links organizational objectives to performance gap identification to the final evaluation of the capacity-building programmes that strive to bridge the gaps. The process starts with organizational priorities and objectives and what is happening in the police organization that indicates a need to change individual or group performance. It then moves to identifying a gap between the desired performance and the actual performance, then investigating whether capacity-building is an appropriate and viable solution to bridging that gap. If the answer is yes, the process continues by identifying the relevant skills, knowledge and character traits that require development and then moves to specifying the learning objectives that will address those development needs.

The evaluation process should mirror the steps in the systematic approach to capacity-building by considering whether the specific performance gaps have been identified, whether the gaps have been bridged and whether this has assisted in achieving the organization’s priorities and objectives. Thus the success of the evaluation phase is directly contingent upon a sound performance gap identification process and perceived difficulties in the evaluation phase are often the result of a badly executed performance gap identification process. Ideally the evaluation phase should be planned at the same time as the performance gap identification in any major programme.

However, the cause of the failure might lie elsewhere. A performance gap identification process must take account of any factor that is linked to the achievement of organizational objectives – this will include factors that relate to organizational structures and policies. The latter cannot be addressed by capacity-building interventions but must be brought to the attention of the appropriate senior managers. Going back to the example of the rights of detained persons, it might be
the case that the lack of infrastructure and resources make it impossible for officers to respect certain rights (e.g. insufficient detention space, inadequate heating or ventilation, insufficient interpreters, a lack of regular meals, no means of contacting family members, insufficient resources to fund defence lawyers at public expense).

Another common problem in police reform programmes lies in the frequent mismatch of capacity-building intervention and workplace application. This can occur in one of several ways. Firstly, officers are often selected for capacity-building programmes that are not immediately relevant to their operational or administrative function, due to a failure or confusion in the selection criteria. Alternatively, problems occur where the correct selection criteria are applied and the right officers attend the capacity-building programme but there are no subsequent opportunities to use or practise the new skills in the workplace.

The nature of evaluation
Evaluation can be defined as a systematic process of measuring the effects of capacity-building. There are a number of types of evaluation, varying in form, the data they yield, and the situation to which they can be effectively applied. The various approaches in common use can be categorised as either empirical or statistical. The latter involves collecting and analysing numerical data from which predictions are drawn and usually involve measuring a representative sample of the entire target group. Whereas, an empirical approach requires the collection of data derived from experience, observations and experiments. In addition to this distinction, most capacity-building specialists will identify the data obtained from an evaluation as either qualitative or quantitative.

<table>
<thead>
<tr>
<th>Chart 9.2: Linking an evaluation to the performance gap identification process</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bridging Performance Gaps</td>
</tr>
<tr>
<td>Devising job profiles that serve organizational objectives</td>
</tr>
<tr>
<td>Identifying poor performance areas</td>
</tr>
<tr>
<td>Capacity-building specification</td>
</tr>
<tr>
<td>Performance gap</td>
</tr>
<tr>
<td>Required skills, knowledge and character traits</td>
</tr>
<tr>
<td>Workplace performance</td>
</tr>
</tbody>
</table>
Evaluation should be used flexibly in meeting the needs of different parts of the police organization. Data derived from an evaluation process can be used to measure the value added by a capacity-building programme, to analyse processes, to prove cause and effect and to acquire diagnostic data for organizational development. It is important to appreciate the broad sweep of application in order to select the most appropriate type of evaluation to meet the requirements of the police organization. The various types of evaluation can be viewed on a scale that runs from the empirical or objective to the subjective or non-empirical approach at the opposite end.

Once a comprehensive police reform process has begun, a number of capacity-building programmes will commence that differ significantly in the knowledge, skills and character traits they seek to achieve: this will, in turn, demand a number of types of evaluation in conjunction with various measurements. The specialist must be prepared to consider programmes on a case by case basis. Some capacity-building programmes will need to be assessed on an individual basis. It may be appropriate in some cases for the specialist to employ a combination of methods and select the most appropriate tools and instruments by which the desired data can be collected.

MODELS OF EVALUATION

There are a number of evaluation models available to capacity-building specialists and the range of choice may appear daunting. However, it is suggested that organizations should look at one or two traditional and proven models that can, if necessary, be adapted to meet the specific needs of the evaluation task.

Kirkpatrick Model
In the context of police reform, a good starting point is the evaluation model devised by D. Kirkpatrick37. This approach, known as the Kirkpatrick Model, has been used in a number of different ways by various organizations, either in an adapted or original form. This model envisages data collection at four distinct levels, as indicated in Chart 9.3.

<table>
<thead>
<tr>
<th>Chart 9.3: Kirkpatrick’s evaluation model</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Advanced level of evaluation</strong></td>
</tr>
<tr>
<td>Results level</td>
</tr>
<tr>
<td>Performance level</td>
</tr>
<tr>
<td><strong>Basic level of evaluation</strong></td>
</tr>
<tr>
<td>Training level</td>
</tr>
<tr>
<td>Reaction level</td>
</tr>
</tbody>
</table>

The lower levels (reaction and training) have an importance for those involved in the design and development of training and development materials, as well as instructor

development. The first level of evaluation (the reaction level) can provide invaluable data on problems that have arisen during the capacity-building programme itself and, sometimes, an insight into the causes if the programme is less than fully effective. Whereas the training level seeks to measure whether the required knowledge and skills contained in the programme objectives have actually been learned. This is usually achieved through a formal test or assessment that employs objective and quantifiable measurements.

The reaction and learning levels are relatively easy to organise but they do not provide any significant indicators of the final test of a capacity-building programme, viewed within the context of police reform: namely, the real impact on workplace performance. Accordingly, Kirkpatrick inserts two further levels. The performance level tries to measure police job performance through a range of evaluation tools over a period of time. Closely allied to this is the results level that seeks to measure the effect that the capacity-building programme has had on the overall performance of the police organization. As you may imagine, this is not easily achieved due to the difficulty in controlling the numerous variables that influence known performance results. However, too often the use evaluation in police organizations remains impotently locked into the immediate reaction level (i.e. through use of so-called ‘happy sheets). In a police reform process the emphasis must be on using evaluation to verify the impact on organizational objectives (the results level) or at least to identify some significant change in the performance/attitudes of the relevant staff (job performance level).

The power of the Kirkpatrick model, therefore, lies in its potential as a diagnostic tool in monitoring progress in overall reform objectives. For example, where an organization has designed and implemented a programme to address the failure of officers to respect the rights of detained persons under the local law, the last two levels might well be ascertained through interviews with arrested persons. If poor results are detected at the job performance level (i.e. a significant percentage of arrested persons report that they were not informed of their rights by trained officers), then performance of all or some of the officers has failed to change in the desired way. Exploration of the lower levels (the programme itself and the instructors) is the first step in isolating the cause of failure in meeting the organizational need through capacity-building. The analysis of evaluation results will be discussed in greater detail in Chapter 11.

**CIRO Model**

Another evaluation approach that lends itself to adaptation is that described in the work of P. Warr, M. Bird and N. Rackham. This approach, known by the acronym ‘CIRO’, is also based on four measurement categories but differs from the Kirkpatrick model in several respects. It envisages four categories of data capture:

- Context evaluation
- Input evaluation
- Reaction evaluation

---

• Outcome evaluation.

As the name suggests, a context evaluation seeks to measure the context within which a capacity-building programme takes place. It scrutinises the way performance needs were identified, learning objectives were established, and the way the objectives link to and support the necessary competencies. In addition, it ought to consider how these components of the programme reflect the culture and structure of the organization. This type of evaluation confirms or otherwise if capacity-building is required. Input evaluation tries to measure a number of inputs to a capacity-building programme, with a view to assisting managers in the process of identifying those which will be most cost-effective. To that end, it focuses on the resources needed to meet performance needs (e.g. staff, facilities, equipment, catering, budget), the content and delivery methods that allow the capacity-building to be achieved, the participating officers, and the results from previous programmes that are similar.

As in the Kirkpatrick model, the reaction evaluation tries to measure how the trainee officers reacted to the programme. Against what was intended by the programme, this type of evaluation draws on the subjective opinions of participants about the capacity-building and how it might be improved. Finally, the outcome evaluation should measure the training and development outcomes against the benchmark of the programme’s objectives. The authors differentiate four levels of outcome evaluation that have strong parallels with the Kirkpatrick model: the learning outcomes of trainees (i.e. changes in their knowledge and skills), the outcomes in the workplace (i.e. changes in actual job performance), outcomes for the relevant areas of the organization (i.e. departments or specialist units), and finally, the outcomes for the organization as a whole.

As in the Kirkpatrick model it is the last of these outcome measures that represents the greatest challenge because of the demand of proving that the capacity-building, as opposed to other factors, effected tangible changes in a police organization. Of course much depends on the nature of the learning objectives. Those that result in tangible, observable and measurable outcomes, such as a reduction of operating costs (e.g. reduced fuel costs for police vehicles), an increase in police services (e.g. crime prevention guidance) and improved work efficiency (e.g. structured patrol methods) will readily lend themselves to this approach to evaluation.

**CRITICAL ISSUES IN EVALUATION DESIGN**

A critical issue in data collection is the existence of a carefully crafted plan, guiding and informing the collection process. A plan will articulate the purpose and objectives of the data collection, as well as the nature and form of the required information. Without these guiding principles the exercise risks being diverted into irrelevant areas or overburdened with useless information. The plan will also set out the design of the study, the methods of data collection and address the difficult matters of data reliability and validity. These three steps require further elaboration.
**Evaluation design** – A first step in the design process is the identification of the type of measurement that will be necessary: in other words, what are we trying to measure? The answer will be guided by the chosen evaluation method and the level of evaluation within that method. The specialist must consider whether quantitative measurements or qualitative measurements or a combination of both are needed to evaluate the quality of the capacity-building provision. All of this must fit with the practical considerations of the police working environment and the nature of the capacity-building programme under consideration. A prudent specialist will reflect hard on the fact that, in spite of the apparently best design, it is not possible to fully isolate or control those external and internal factors that lie outside the scope of capacity-building interventions. Reality dictates that - as in police work generally - the specialist must seek to achieve a compromise, balancing time and resources against the need to obtain reliable and meaningful results.

Next the evaluation design ought to clearly establish the target group, the use of a sample or a census, and the experimental design that will be employed. The first issue – the target group – relates to those from whom the specialist hopes to get information. This might include the trainee officers from some or all of the capacity-building programmes, their supervisors or those whom they supervise, their peers, or their instructors, or the external recipients of the relevant police service. Whilst the trainees might seem the most obvious choice, an objective evaluation ought to seek information about performance from those who might be classed as the ‘beneficiaries’ of the relevant police services. Of course the instructors will play a very special role as sources of good data on how well the capacity-building programme met its objectives and the efficacy of the learning methods that were used.

**Quantitative measurement** - Based on an empirical and scientific approach, quantitative measurement is normally inspired by the principles of hypothesis, objectivity and deduction. There is a natural inclination among police managers to prefer this - the ‘number crunching’ approach - to the more qualitative measures. This prejudice tends to overlook the desirability for a careful mixture of qualitative and quantitative measurement. There are areas of the policing workplace where it is simply not possible to prove cause and effect and a qualitative approach may be more appropriate. The capacity-building specialist must carefully consider the needs of each situation, deciding whether a quantitative measurement or a qualitative measurement or a mix of both is the most appropriate path to achieving useful data. This decision involves an assessment of whether the measurement can utilise numerical data or whether the relevant data involves verbal or written information. In the latter case, qualitative measurements are required and in the former case, quantitative measurement will be possible.

**Qualitative measurement** - Qualitative measurement is concerned with subjective and descriptive information, such as the words and the observed behaviour of police officers. Unlike quantitative measurement, it deals with the human dimension rather than reducing performance to statistical formulae because it is based on a firm conviction that all staff and work settings are worthy of measurement, regardless of their preconceived merit. In this sense, it adopts a holistic vision of workplace settings and the staff who operate in them. This does not, however, preclude accuracy and validity, even though reliability might be ignored. Using an inductive approach, qualitative measurement allows the
measurement process to develop concepts and understanding on the basis of patterns that emerge from the data rather than from preconceived models or hypotheses about those patterns. Most importantly, a specialist must be aware that the patterns that emerge from the data cannot rely on his or her beliefs and perceptions. Rather, the specialist must stand aside and allow the data to measure reality as the police officers see it, thereby providing an insight into police performance within the frame of reference of the reflections and observations of the staff.

Scope of data capture - Our next concern relates to the scope of the evaluation exercise: should it address every member of the target group (a ‘census’) or merely a subgroup of that group (a ‘sample’). Obviously sampling will be the favoured approach when evaluating large-scale capacity-building programmes, such as training and development in new legislation for all police staff or major sections of the organization. A census becomes very costly and time-consuming if the target group runs to hundreds or thousands of staff. There are several different methods of sampling and the use of sample results is a complex matter that requires a significant degree of expertise. Where a sample is chosen randomly it can be possible to draw valid conclusions about the entire target group.

Control mechanisms - Just as pharmaceutical companies must trial new drugs in order to test their efficacy and side-effects, it is important that capacity-building programmes are tested to see whether they achieve what they are supposed to do. There are two basic experimental design approaches that involve only the participants and can be readily adopted for a particular evaluation study. Firstly, one can take measurements after the capacity-building programme, making comparisons against an agreed yardstick of the desired performance. Whilst this approach allows for checking that the programme has achieved its objectives, there is no way of assessing how much the capacity-building intervention has contributed to the achievement of these objectives without some measurement prior to the intervention. Without some other evidence to the contrary, it is possible that the participants might have achieved similar results without ever attending the capacity-building course.

The alternative involves measuring performance before and after the programme, thereby assessing any significant gains. However, it is possible to argue that the trainee officers would have achieved the measured gain or a significant part of it without the programme through workplace experience, if the capacity-building programme was spread over a period of time. In spite of these potential shortcomings, both methods are widely used in the evaluation of police programmes and provide useful indicators for capacity-building specialists. There are of course more sophisticated design approaches that involve the use of control groups and their merits ought to be considered briefly.

Use of control groups - As the name suggests, control groups are used in an effort to obviate the effects of extraneous factors that might influence the results of an evaluation. The idea is cunningly simple. First the specialist selects out police employees who are as similar as possible to the target group (i.e. the trainees) to form a control group, then assesses the performance of both groups prior to the start of the programme. While the target group receives the capacity-building, the control group does not. Finally, both groups are again assessed after the programme and the gain in performance of the target group is compared with the gain in the control group’s performance. The impact of the programme, over and above extraneous factors, should be revealed in the difference in performance gain.
The use of control groups is not without problems. In practice it may be difficult or impracticable to find control groups among police staff that are similar to the target group. There is an argument that the principle behind the use of control groups is undermined by the fact that merely experiencing a capacity-building programme will influence performance (the Hawthorne effect). The majority of evaluation studies in police programmes never use control groups and it is arguable that, in the early period of a reform process, the difference between desired performance and actual performance is great enough to allow the use of less rigorous measurement techniques. In a reform setting it will often suffice if the specialist measures performance after the programme against the yardstick defined in the relevant job profile.

METHODS OF DATA COLLECTION

The main methods of collecting data are identical to those methods of investigating the performance gap: questionnaires, interviews and observation. As was seen in Chapter 6, each method has its disadvantages and advantages and these must be kept in mind when deciding on the most appropriate approach to collecting data in the evaluation process. In this section each method will be considered as it applies to evaluation and useful ways of integrating the data collection tools with the capacity-building programme itself.

Questionnaires

The questionnaire is the most popular method of data collection in evaluation studies. It carries with it important advantages and disadvantages, with a necessary trade-off between the depth of information and breadth of coverage and costs. Chart 9.3 provides a cursory description of these factors. Too often questionnaires are automatically chosen as the method of data collection, without due consideration of their limitations and the benefits of alternative techniques. As in performance gap identification, the best approach may involve a combination of techniques rather than reliance on data through a single channel. Another problem area is that of poor questionnaire design. Much has been written on the complex subject of questionnaire design and it is recommended that specialist skill development is required prior to attempting this task. Within the context of this book, it will assist if some of the main issues in a design process are examined.

Careful attention must be given to the type of information required and the type of questions that will supply this information when designing questionnaires. In addition, thought must be given to how the information is to be processed and analysed, mindful of the considerable costs involved in manual processing. Respondents are more likely to complete forms that are ‘user-friendly’ and, therefore, consideration should be given to the ease of use and the time required to complete the questionnaire.

Experts in questionnaire design normally indicate five main categories or classes of question:
open questions
classification questions
structured questions
differential-type questions
Lickert-type questions.

Each of these types of question will now be discussed in turn.

Open Questions - An open question is one that is so constructed as to give the respondent complete freedom in providing a descriptive answer, as opposed to a closed question that is designed to simply elicit a 'yes' or 'no' answer. Open questions begin with a verb that invites a detailed response: e.g. describe..., tell..., state... For example:

'Describe what you found most useful in the capacity-building programme?'

'Give details of the information that ought to be given under this procedure?'

This type of question is used to test knowledge, particularly in more complex areas or as a means of eliciting information about attitudes. They have the advantage of being easy to design and encourage a free expression of views without the influence of a desired response or other form of bias. On the downside, open questions provide responses that are difficult to analyse and categorise. A strong framework is required to reduce the disparate responses into coherent and meaningful patterns, particularly in the case of attitude and opinion data.

Classification questions - These are questions that are designed to elicit responses that enable a specialist to categorise or classify respondents under particular groups, thereby assisting the process of analysing the response data. Using criteria such as age, gender, ethnic origin, occupation, and rank these questions help to identify the effect of a capacity-building programme on different groups and check how representative the sample is of the entire target group.

An important consideration is where to place classification questions in the questionnaire. In a reform setting this type of question may inhibit some officers from responding because of ethnic or gender issues. Experience indicates that they should be left until the last part of the document. Always ensure that a full range of options is included and, if the questionnaire is being completed anonymously, do not allow the combination of classification questions to unwittingly identify an individual staff member.

Structured questions - Structured or 'coded' questions are popular in a questionnaire format since they limit respondents to a finite choice of answers. They are useful in establishing facts, testing knowledge and measuring attitudes. For example, a question such as, 'How long is a police officer permitted to carry a firearm without undergoing a recertification programme?' will have three or four possible answers and the respondent is invited to tick the box opposite the appropriate response. Whilst these questions are quick and easy to complete and later analyse, they are difficult to design in practice
since it is vital to provide a comprehensive set of options and they are inclined to induce bias in the responses.

**Differential type questions** - These are sometimes called semantic questions because of the stress on a graded and more meaningful response, in order to assess skills and measure attitudes. Each respondent is required to assess an issue on a 7-point scale, for example:

‘Please assess the listening skills of the trainee officer, by circling the appropriate rating.’

<table>
<thead>
<tr>
<th>Strong empathy</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>Weak empathy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listened well</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>Did not listen</td>
</tr>
</tbody>
</table>

Such questions create a highly structured range of responses that are easy to analyse, but invite highly subjective judgements in terms of scale rating (i.e. what might be listed as a 3 to one person, might be a 5 to another person. Moreover, this approach creates a number of complex issues such as what range of scale should be used (4 to 10 is the most common range), whether or not to be consistent with a favoured extreme (i.e. left or right), and whether to employ a middle option (i.e. using an even-number scale to force a choice between the top and the bottom of the range).

**Lickert-type questions** - This is now a common approach to question design in questionnaires. Essentially it involves inviting the respondent to indicate his or her views against a specific statement and is used to measure attitudes and assess skills. For example, a question might seek to assess an officer’s views on a new disciplinary procedure by measuring his/her approval or disapproval of two statements:

- It is easy to understand
- It will improve professional standards in the police service

In order to obtain a graded response to each statement, the officer must select one of the following:

- Strongly agree
- Agree
- Not sure
- Disagree
- Disagree strongly

Lickert-type questions are frequently used and have the advantage of allowing for a structured range of responses. However, as with similarly structured methods, this approach tends to constrain response and induce bias in the results. To offset these disadvantages it is important to create a balanced set of response options (i.e. favourable/unfavourable responses) and give due consideration for a middle option, as in the case of differential-type questions.
Above all it is vital that questionnaires are as brief as possible and contain questions that are as short and simple as possible. Police officers, like other professional groups, have a low tolerance for long questionnaires and tortuous questions. Likewise, question the recent experience of the target group rather than asking about events that they are likely to have forgotten. Try to avoid ambiguous questions and overly technical language or jargon: if the data you seek cannot be translated into a simple question it is probably not worth acquiring. Great thought must be invested in the design of questions. As in any court room, the use of leading questions and emotive words must be avoided (e.g. ‘Do you feel that your supervising officer should be more supportive?’). If you are in search of actual attitudes you should avoid hypothetical questions such as ‘What would you do if a prisoner was shouting and being abusive in police detention?’

<table>
<thead>
<tr>
<th>Advantages</th>
<th>Disadvantages</th>
<th>Advice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Potentially rapid method of collecting data</td>
<td>Potential for low response rates, resulting in biased results</td>
<td>Encourage the return of questionnaires through monitoring and inducements. Ensure that the questionnaire is short and user-friendly.</td>
</tr>
<tr>
<td>Potentially a low-cost method of collecting data</td>
<td>Tends to be unsuitable for capturing detailed information</td>
<td>In the design phase you should closely examine the data required and whether the questionnaire can achieve the desired outcomes.</td>
</tr>
<tr>
<td>Imposes few demands on the participants</td>
<td>Demands careful design to ensure clarity and the absence of ambiguity</td>
<td>Consider hiring a consultant or procuring specialist capacity-building for evaluation team.</td>
</tr>
<tr>
<td>Takes less time than alternative methods</td>
<td>Potentially unsuited to certain target groups that do not normally complete forms</td>
<td>Invest in additional capacity-building/briefings on the completion of forms.</td>
</tr>
<tr>
<td>Provides results that can be readily analysed if carefully designed</td>
<td>Can breed contempt if used too often or used in a way that does not allow police officers to see the benefit in completing the forms</td>
<td>Ensure that you subject the questionnaire to a pilot test, thereby checking the quality of the data it can obtain.</td>
</tr>
</tbody>
</table>
Interviews
There are of course various ways of gathering evaluation data through interviews: the individual (or ‘one-to-one’) interview, focus groups, telephone interviews and – in more recent times - video-conferencing interviews. Reference should be made to the notes on structured interviews and semi-structured interviews in Chapter 6, since all of these approaches will require preparation and a plan based on some sort of structure. Interviews can range from the highly structured to the relatively unstructured. The latter, the most simplistic approach, might involve the interviewer simply asking the interviewee to talk about the capacity-building event. There will be no pre-planned structuring, other than a conscious effort by the interviewer to guide and focus the discussion toward what is relevant. Somewhere between the structured and unstructured approach stands the partially planned interview design, with guidance notes and even particular questions. Nevertheless, there will be varying degrees of latitude in pursuing and probing replies to questions.

The use of the telephone as a means of interviewing is largely underused in police programme evaluation. It is potentially cheaper in terms of staff abstraction, travel time and transport costs in a police organization that covers a large geographical area. However, it is more impersonal and lacks the power to explore and probe issues in the same way as a traditional ‘one-to-one’ interview. Inevitably telephone interviews follow the pattern of self-complete questionnaires, using tick-box-type responses in a highly structured approach. Although expensive in terms of time for both interviewer and interviewee, the traditional one-to-one interview is a very productive method of gaining useful data. It enables responses on particular issues to be explored and clarified, as well as allowing more complicated matters to be probed. In order to be effective, the individual interview demands that the interviewer is highly skilled and aware of the dangers of interviewer bias (i.e. where the interviewer wittingly or unwittingly influences the responses of the interviewee). Great skill is also required in the analysis of the data from this type of interview, due to the qualitative nature of the data obtained.

As discussed in Chapter 6, interviewing a representative group (a ‘focus group’) is an excellent way of discussing relevant issues, delivering high quality information about the capacity-building event and allowing problem areas to be fully explored. The interviewer must be prepared to exercise close control since focus groups can, in practice, be taken over by one or more strong personalities that block or swamp contrary views within the group. Once the required infrastructure and capacity-building is in place, video-conferencing interviews can reduce the time and expense involved in normal focus group interviews and individual interviews. Once these investments are made it offers significant cost savings in staff abstraction from the workplace, travel time and transport.

Observation
Since police capacity-building is often concerned with observable skills and behaviours, it is an obvious step to adopt direct observation as an important method of data collection in the evaluation of such capacity-building. Of course it can be a costly process, involving a minimum observer/trainee ratio of 1:1, as well as expensive in terms of facilities, time and staff abstraction from the workplace. The design phase must be mindful of the Hawthorne effect where a trainee officer's behaviour is influenced by the simple fact of being observed during a possibly complex task.
being observed during a possibly complex task. This significant factor can be off-set by developing the observers' skills, creating low-impact observation measures and carefully designed observation forms. The notes in Chapter 6 are equally applicable to programme evaluation.

**Structuring and recording data**

As you will have noted, the questionnaire features in all the data collection methods, in one form or another. At the highly structured end, a self-complete questionnaire is a tool that must stand on its own without the need for further explanation that might otherwise be provided by an interviewer or observer. Whereas an interview method necessitates the use of an interview form that will be as highly structured as a self-complete questionnaire or, at the opposite extreme, a document that merely establishes the main subject headings that will be addressed.

If an observation method is adopted, there will be a need to record the observed behaviour on an observation form that might include closed questions (requiring a yes or no response) about the presence or absence of a specific behaviour or skill and its frequency, or a form that grades observed skills and allows subjective comments. All questionnaires need to be well designed and those used in more formal surveys should undergo an informal trial on a small group that is similar to the target population. The results of the trial can be used, as necessary, in redesigning the form(s). It is useful to consider the inclusion of classification questions that can gauge the way the effects of the capacity-building vary from one group to another.

**VALIDITY, RELIABILITY AND RELEVANCE**

Important decisions may be based on the results of an evaluation and it is important that they are able to demonstrate that they meet a defined standard of transparency. This will entail an ability to stand up against a good degree of scrutiny in the areas of internal validity, external validity, reliability and relevance to reform objectives.

**Internal validity** - As the name suggests, internal validity is concerned with measuring performance against internal factors: in other words, it is concerned with how well the evaluation measures what is desired as an outcome or what it was aiming to find out. This directly relates to the instrument used to collect information and its adequacy and appropriateness as a measuring tool. Does the questionnaire that was used contain questions that are appropriately worded to extract the required data? Does the knowledge check measure the knowledge that must be learned in an effective and helpful manner? Does the skills test contain a sufficient profile with measurement scales that are readily understood and appropriate? The process of establishing the internal validity of an evaluation is assisted through testing instruments and using alternative approaches to measure the same attribute.

**External validity** - If an evaluation study has measured the effectiveness of a particular capacity-building programme by looking at 10 trainee officers from a total group of 50, the question arises as to whether the results can be applied to the entire group. This important question relates to the external validity of an evaluation study -
the extent to which the findings can be applied beyond the group used for the purpose of the study. A precise answer involves entering a complex study of sampling, taking us beyond the scope and purpose of this book. However, an affirmative answer can be obtained if two factors are accepted as present. Firstly, the sample of the 10 trainees must be selected appropriately and thereby shown to represent the entire group. Secondly, if it is acknowledged that an evaluation result is merely an estimate for the whole target group and that the result for the whole target group lies somewhere in a range around that estimate, then the answer must be in the affirmative.

A more relevant question in a police reform programme is whether a capacity-building programme that is assessed as effective in one region or area of the organization can be declared effective for the entire organization. Given the higher number of similarities among police trainees this sort of generalisation from a limited sample is more likely to be valid than in other organizations in the public or private sector. The validity of such a generalisation must be based on known similarities, in terms of the experience of the organization and start levels of knowledge and skills among trainee officers. Whilst it is important that you strive to demonstrate the wider validity of an evaluation, it is important to be very clear about the assumptions that underpin your statements and exercise care in what you claim to be the successful outcomes.

Reliability of the results - The next point of interest relates to the reliability of an evaluation: in other words, the extent to which its results can be replicated. If the evaluation study was repeated, would the results be identical or very similar? A number of techniques can be applied, including the same question in a questionnaire but in different forms, or using multiple observers, or simply repeating tests and observations.

Relevance to reform objectives - Some may well question the value of evaluation, given the cost and effort: surely in a world of limited resources the money would be better spent on capacity-building itself? Likewise, others are understandably perplexed by the list of requirements for the design of a sound evaluation study. In practice it may not be possible to achieve such rigorous standards in every evaluation. The specialist must be aware of the issues and solutions that will allow the achievement of a professional standard in evaluation work. The focus should remain on the reform-related question that underpins the evaluation and then invest effort in optimising the quality and integrity of the data, since there is little point in acquiring accurate data that fails to yield useful information.

An appropriate design phase means bringing together data about the officers' profiles, alongside data about how they learn, the nature of the required knowledge and skills and the desired outcomes of the programme. A capacity-building programme will prove inevitably inappropriate if it is designed without taking into account the relevant reform objectives, officers' extant competences, background, work experience, education, culture and language (where more than one language is used in the organization). Likewise the programme should reflect the way the officers learn, whether with strong links to the workplace or through an emphasis on previous experience. Where a programme proves ineffective, an analysis of this area might provide useful indications of those matters that require urgent attention in relation to the programme design. In this way the value of programme evaluation will become self-evident as the pivotal indicator of success or failure in achieving reform objectives through capacity-building.