ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT

FORUM ON TAX ADMINISTRATION: SMALL/MEDIUM ENTERPRISE (SME) COMPLIANCE SUB-GROUP

Companion Guide

Methdological techniques for use in evaluating the effectiveness of compliance risk treatments

November 2010

CENTRE FOR TAX POLICY AND ADMINISTRATION

Table of contents

Pages

Annex

ABOUT THIS DOCUMENT

Purpose

This companion note provides refernece material that may assist member revenue bodies in planning their evaluation of compliance risk treatment strategies. It should be read in conjunction with the Forum's guidance note—*Evaluating the Effectiveness of Compliance Risk Treatment Strategies.*

Background to the Forum on Tax Administration

The Forum on Tax Administration (FTA) was created by the Committee on Fiscal Affairs (CFA) in July 2002. Since then the FTA has grown to become a unique forum on tax administration for the heads of revenue bodies and their teams from OECD and selected non-OECD countries.

In 2009, participating countries developed the *FTA vision* setting out that… *The FTA vision is to create a forum through which tax administrators can identify, discuss and influence relevant global trends and develop new ideas to enhance tax administration around the world***.** This vision is underpinned by the FTA's key aim which is to…*improve taxpayer services and tax compliance – by helping revenue bodies increase the efficiency, effectiveness and fairness of tax administration and reduce the costs of compliance.*

To help carry out its mandate, the FTA is directly supported by two specialist Sub-groups—Compliance and Taxpayer Services—that each carry out a program of work agreed by members. Both OECD and selected non-OECD countries participate in the work of the FTA and its Sub-groups.

The Compliance Sub-group's mandate, in broad terms, is to provide a forum for members to:

- periodically monitor and report on trends in compliance approaches, strategies and activities;
- consider and compare member compliance objectives, the strategies to achieve those objectives and the underlying behavioural compliance models and assumptions being used;
- consider and compare member compliance structures, systems and management, and staff skills and training; and
- develop and maintain papers describing good country practices as well as develop discussion papers on emerging trends and innovative approaches.

Since its inception, the Sub-group has focused its work on issues associated with improving the tax compliance of SME taxpayers.

Caveat

National revenue bodies face a varied environment within which to administer their taxation system. Jurisdictions differ in respect of their policy and legislative environment and their administrative practices and culture. As such, a standard approach to tax administration may be neither practical nor desirable in a particular instance. The documents forming the OECD tax guidance series need to be interpreted with this in mind. Care should always be taken when considering a country's practices to fully appreciate the complex factors that have shaped a particular approach.

Inquiries and further information

Inquiries concerning any matters raised in this note should be directed to Sean Moriarty (Head, International Co-operation and Tax Administration Division) at e-mail (sean.moriarty@oecd.org).

I. Methdological techniques for use in evaluating the effectiveness of compliance risk treatments

Background

1. At meetings of the Compliance Sub-group in both 2008 and 2009, members acknowledged that there was a critical gap in the detailed practical guidance available to revenue bodies for fully implementing the recommended risk management process. The subject of this gap was the practical approaches and methods that could be used for systematically evaluating the effectiveness of specific compliance risk treatment strategies. However, it was acknowledged that work was underway by both the Australian Taxation Office (ATO)[1] and by European Commission (EC) (as part of its Fiscalis program [2]) to develop practical guidance in this area which might serve as valuable input to guidance that the sub-group could prepare. Members accordingly agreed to initiate work to develop a set of practical guidance in this field.

2. In giving direction for this work members requested that the guidance should build on work already done in this field, limited as it is, and not aim for absolute precision, recognising that evaluation in the field of taxpayers' compliance was 'more of an art than a science'. Furthermore, it should encompass ideas for its practical implementation in an organisational sense and, in particular, should: 1) be oriented towards senior managers (as opposed to technicians); 2) be practical and not too academic; 3) have a clear 'outcomes' orientation; 4) provide an overview of measurement approaches that are feasible; and 5) be supported by good case study examples to demonstrate the recommended techniques.

Introduction

3. The guidance note now prepared describes a formal, structured and systematic process for conducting evaluations of specific compliance risk treatments. In doing so, it illustrates how various statistical and other methodologies can be deployed to help understand the impacts of specific risk treatments. This compendium provides access to an overview of some of the more common methodologies, along with a description of their recognized strengths and weaknesses, that can be used in this domain that may be of interest to revenue bodies and their staff.

Acknowledgement

4. The reference material contained in this guide—see Annex 1— draws extensively on the document—*An Overview of Evaluation Methodology and Techniques*—prepared by the Fiscalis Risk Management Platform Group and published by the European Commission (EC). Some additional material was found in research findings reported in an ATO publication—*Literature Review: Measuring Compliance Effectiveness*. Forum members are indebted to the EC's Fiscalis program for their initiative in this field.

[1] The ATO published its practical guidance material on methodologies for measuring the effectiveness of its compliance strategies commenced in August 2008. Copies of these can be found at: http://www.ato.gov.au/complianceeffectiveness

[2] Officials working as part of the European Commission's Fiscalis program included guidance on evaluation as part of a Compliance Risk Management Guide for Tax Administrations, a revised edition of which was published in March 2010 —see http://ec.europa.eu/taxation_customs/resources/documents/common/publications/info_docs/taxation/risk_managt_guide_en.pdf

Annex 1

Methodological techniques for use in evaluation

Experimental

Randomised controlled trials (RCTs)

The randomised controlled trial is the ideal evaluation methodology to infer program effectiveness. It is the only method that can eliminate the influence of other known and unknown factors, which can produce misleading evaluation results. The trial compares a control group and treatment group and assumes the differences are due to the activity being evaluated, as illustrated in figure below:

The control group should be analytically similar to the treatment group (that is, representative) and of an appropriate size to enable statistical tests to be performed. Groups should be randomly selected *at the beginning* of the program, so program administrators must factor in evaluation needs.

Random selection reduces systematic differences in any characteristics (such as demographic factors) between the groups, leading to increased confidence that any differences can be attributed to the intervention. Generally, trials can be undertaken where:

- program participants and non-participants can be randomly assigned to two or more groups large enough to comprise a statistically valid sample;
- each group can be administered a distinct intervention, or non-intervention (for the control group), and
- the outcomes the intervention was designed to improve can be measured for each group.

To ensure that a trial is well designed and implemented, the study should clearly describe:

- the intervention, including who administered it, who received it, and what it cost
- how the intervention differed from the treatment the control group received, and
- the logic of how the intervention is supposed to affect outcomes

Strengths

- Clear and easy to explain to non-technical audiences.
- Randomisation ensures no systematic differences on average between treatment and control groups. Hence, the only two differences between intervention and control group are the impact of the programme and random differences.
- Do not require complex statistical techniques.
- Can be a fair mechanism of allocating interventions where resources and need to be rationed

Weaknesses

- Not always ethical. It can be perceived that some people are being denied entry to the programme.
- The trial may be biased as those deterred from participating may be sub-groups who are the most risk-adverse, or may perceive the potential impacts to them as negative or who have less understanding of the issues.
- Findings and causal relationships detected under a RCT may only be true for the geographical or target areas that they were tested in.
- May be difficult to carry out or it may not be possible to randomly allocate.
- Can be expensive i.e. programmes have to be run within areas at the same time.

Quasi–experimental

Matched area comparison (pre-matching)

Treatment tested in a number of areas in the country with and then compared to a matched area with similar characteristics where treatment is not tested.

Strengths

- Usually more ethically acceptable than randomisation since all in the same area are treated alike. Can still be some objections in treating areas differently.
- Appropriate when the expected impact of the programme is considerably greater than any variation that might be expected between the areas if the intervention was not in place

Weaknesses

- Difficult to separate out the difference from the programme to the differences in characteristics the treatment and comparison groups, systematic differences between the areas and differences induced by other programmes.
- Currently no common consensus exists on criteria for matching pilot areas with comparison areas. One way is to match areas that had similar outcomes before the intervention, but this relies on historical data being available. Other ways are to match on a set of economic and social indicators. Sample sizes, competing initiatives and willingness to take part may limit what areas can be included.
- Although areas may be matched well at time of selection, they may have drifted apart by the time outcome data is collected.
- If evaluators handpick areas they may be accused of picking those that are likely to ensure a good measure of additionally is observed. This is more of a risk if the evaluator is not independent from the programme.

Non-experimental

Matched comparison group (post-matching)

Participants and non-participants groups select themselves or by selection from case-handler (in control activities for example). Evaluator selects an intervention group from the population of participants and a comparison group from the population of non-participants, i.e. matching is done post treatment. Statistical methods are used to match each selected participant uniquely to a non-participant. Non-participant comparators should be identical to their matched counterparts in all factors associated with participation and the outcomes of interest with the exception that participants participate and controls do not. Groups are followed over time and the impact of the programme is the difference between the two groups.

Two common methods of matching are ‘cell matching’ and ‘propensity score matching’. Another matching method is ‘genetic matching’ that, according to its developers, provides estimates that are more reliable and more robust to small changes in data. One suitable input variable to the matching algorithm is a known or estimated propensity score. The method can therefore be considered a generalisation of propensity score matching

Strengths

- Can be useful for evaluating the impact of voluntary implemented policies.
- Since no-one is denied access poses few ethical difficulties.
- Reasonably robust estimate of the additional impact of the programme where all factors affecting participation are known and data on these factors can be collected

Weaknesses

- Robustness dependent on the matching procedure. Poor matching can introduce bias e.g. if important variable are omitted from the matching criteria. Hidden bias may also remain because matching only controls for observable and available factors.
- Matching is only as good as the data and statistical technique used. Where information cannot be collected retrospectively, the information will need to be collected prior to participation.
- Large samples are required to create sufficient matches (although these may be smaller than under the matched area comparison design).
- Data collection can be expensive.

Non- experimental

Simple before and after Studies

In a before and after comparison, progress is measured at two points; before and after. The difference between the two measurements is the impact of the programme/policy.

Strengths.

- Give reasonably robust estimate of the additional impact of the programme when the impact is large compared to changes from external factors.
- Can be combined with matched comparison designs or matched area designs to increase robustness and factor out historical change and observed differences in factors between the groups. Difference-in-Differences (DID), ANOVA (Analysis of Variance) or Non-Parametric Tests and Regression Discontinuity Design are other types of more robust before and after design.

Weaknesses

- Difficult to separate out change attributable to the programme from general historical change i.e. natural change that would have occurred anyway.

Natural experiment

Instrumental variable method

In some cases it is not possible to use a matching algorithm on the basis of observable variables. In those cases it could be possible to identify a so called instrument, that is a variable that fulfils two criteria: (1) the instrument at least partly explains assignment to treatment (who belongs to treatment group and not), but (2) does not directly determine the impact of the treatment. It is a search for a natural experiment, one can say. By studying how the outcome varies for different values for the instrument variable, an estimate of the impact can be calculated

Strengths

- When the criteria are fulfilled, estimates of the impact are robust.
- The method does not need to influence the compliance activity, but the intervention is targeted to those with the greatest need or those that yield the greatest benefits or likelihood from selection of meeting objectives.

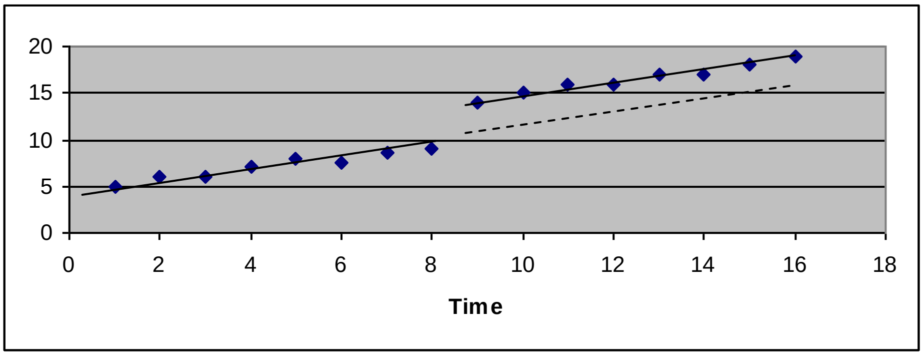
Weaknesses

- Not possible to empirically test the validity of the choice of the instrumental variable.
- Can be hard or impossible to find an instrument variable of good quality.

Non-experimental

Interrupted time series (ITS)

ITS monitors repeated observations over time from before the programme/treatment and after. An extension of the simple before and after approach where there is no control group. A break in the time series trend at or shortly after the treatment can be interpreted as the impact.



Strengths

- Extending the number of before and after periods can sometimes overcome the problem of isolating programme change from natural change.
- If no related treatments are introduced around the same time, then a sudden change in the series could be fairly conclusive. Especially true if change observed is larger than the change observed between any two of the before periods. If the change is sustained over time then the evidence will be even stronger

Weaknesses

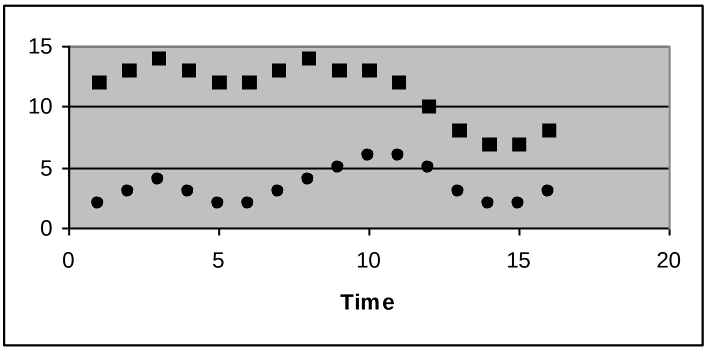
- Some interventions have a delayed or gradual impact so interruption in the time series could occur some time after implementation.
- If the introduction of the programme coincides with other events/programmes that have an impact on the same outcomes, it will be difficult to determine causality of the outcomes observed and how much is due to the intervention being evaluated.
- Smaller policies may be difficult to separate out their impact from background noise or natural variation in the time trend.
- Usually only possible to use administrative data or standard datasets.

Non-experiment

Difference-in-differences (DID)

DID compares the difference in average outcomes of the intervention group with the comparison group for at least one relevant point before and after implementation. Efforts are made to find a comparison group which had little or no exposure to the treatment. The power of DID increases with the number of before and after point estimates examined. In essence, DID “differences out” the differences between both groups and the common time effect.

In the example below, the difference-in-difference between the two groups is used as an estimated of the effect. The action is started at time 8 and has reached full impact at time 12. The effect is estimated to 5 units, as the difference between the two groups has decreased from 10 to 5 points.



Strengths

- Reduces the risk of measuring historical change or change due to factors other than the programme, which basic before and after designs and interrupted time series can suffer problems from.
- Groups can start from different levels as this approach focuses on change rather than levels

Weaknesses

- DID (unlike ANOVA) assumes that there is no variance between groups which may not be the case. So although DID may be a more robust method than some other designs in controlling for change not attributable to the intervention, as some of the Examples below show in practice DID cannot or does not always completely control for these.

Analysis of Variance (ANOVA)

(ANalysis Of VAriance) is to test differences in means (for groups or variables) for statistical significance. It is based on comparing the variance between the means for sample groups with the variance common to all groups within the analyses. ANOVA is therefore carried out when groups are expected to be different e.g. there is sampling variability and tests the null hypothesis that the treatment means in the population are equal. ANOVA assumes that: Observations within each sample are independent and samples are normally distributed.

Strengths

- Unlike the Difference-in-Differences (DID) approach, ANOVA takes into account sampling variability.
- ANOVA can test whether several means differ.7 ANOVA therefore tests all means at once, by comparing everything in a hierarchical manner to one fixed point. In this way it controls for the type 1 error that would result by performing *t*-tests between each of the groups, ('Type I error' = rejection of the null hypothesis when it's correct e.g. where assuming a 95% confidence level one in every 20 *t*-tests would not be valid due to type 1 error).
- It is not affected by multi-collinearity. It allows us to deal with two or more independent variables, testing not only the individual effect of each variable separately but also the interacting effects of two or more variables

Weaknesses

- ANOVA requires large sample sizes and involves a lot of complexity especially when compared to the Difference-in- Differences (DID) approach. Unless sample variability is an issue (e.g. control and intervention groups are likely to be different) it will have little to add to techniques such as DID.

Natural experiment

Regression discontinuity design (RDD)

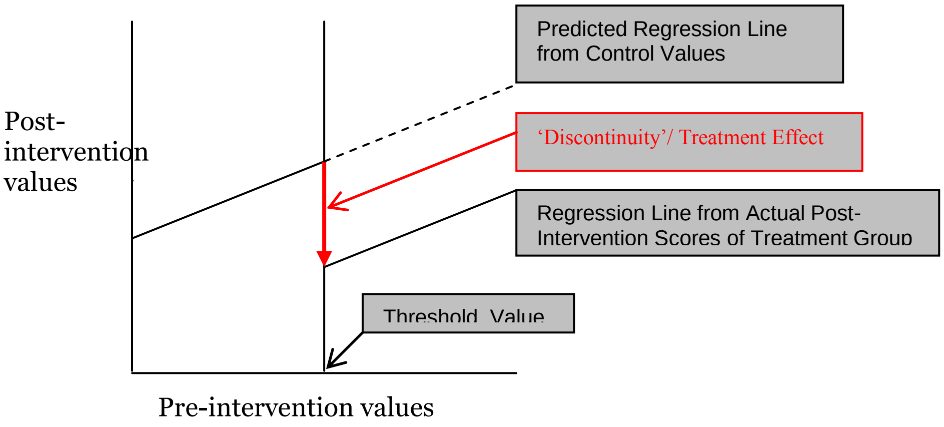
Regression Discontinuity Design (RDD) is a fairly new technique and has been acknowledged as the quasi-experiment method that comes closest to RCTs in eliminating selection bias. RDD is a before and after design which involves splitting participants from the eligible population into programme and comparison groups based on whether they are above or below a known threshold (or cut off point/assignment score/eligibility criteria) of an observed continuous assignment variable. In tax policy, the assignment variable will probably most commonly be income. All participants are assessed before the intervention and reassessed after receiving the intervention. Each point on the graph to the right shows the relationship between the pre and post measures for each person in the study. A regression line is drawn between all the observations for participants who did not receive the intervention (e.g. those on the left hand-side of the cut-off point). This regression line then projects the post-intervention scores for the cases who receive the intervention. These predictions are compared to the actual post-intervention values for those who received the programme. The *discontinuity* or observed jump between the predicted and actual post-test scores is taken to be the 'treatment effect' of the programme.

The discontinuity can also be found with the help of qualitative factors, or combined with quantitative factors, when sharp or fuzzy criteria are used by the administration to assign intervention. For example, an information activity is assigned to businesses aged 1-3 years old, located in two specific regions and with a VAT return between X and Y euro. Relevant comparison groups can be found by investigating discontinuities in the assignment variables age, region and amount of VAT return, one at the time.

Strengths

- Unlike with other quasi-experimental designs, the source of selection bias is known and quantified e.g. the only source of selection bias in RDD is the assignment variable. In other forms of quasi-experimental designs commonly only a limited number of variables are controlled for, but an infinite number of unknown variables could influence the result of the evaluation. It is thus the method that comes closest to RCTs in eliminating selection bias.
- Since participants are assigned to groups on a cut-off score, the intervention is targeted to those with the greatest need or those that yield the greatest benefits or likelihood from selection of meeting objectives. This is in contrast to RCTs where allocation is random

Figure


Predicted Regression Line from Control Values
'Discontinuity'/ Treatment Effect
Regression Line from Actual Post-Intervention Scores of Treatment Group
Threshold Value
Post-intervention values
Pre-intervention values

Weaknesses

- Large sample sizes are needed in order to deduce if treatment effects are significant. Data needs to be collected both before and after the intervention.
- Statistical analysis can be complex.
- RDD can only be used where there is an appropriate continuous assignment variable.

Non-Parametric Tests

'Non-parametric' tests are typically described as 'distribution free'. Often used in place of their parametric counterparts when certain assumptions about the underlying population are questionable. Non-Parametric tests may be, and often are, more powerful in detecting population differences when certain assumptions are not satisfied.

Strengths

- Simple to employ and are based on far more liberal assumptions than their parametric equivalent. Hence they can be used far more widely and flexibly.

Weaknesses

- To achieve the same level of certainty, generally a slightly larger sample is required for a non-parametric test than its parametric equivalent.

Natural experiment

Structural models

Structural models are constructed from micro-data and use equations to model the factors that influence demand and supply within a particular sector of the economy. A structural model can then be used to simulate the impact of a programme/policy separately from the impact of other factors.

Strengths

- Enable the economic mechanisms leading to identified effects to be separated out, e.g. it allows the impact of changes caused by the intervention to be disentangled from other factors.
- They allow examination of the distribution of individual effects.
- They can be a useful source of estimating the impact of the proposed counterfactual and in forecasting the dynamic effects of an intervention before it is implemented.

Weaknesses

- Time-consuming and complex to build and are based upon strong assumptions. A lot of structure is required to build a believable behavioural model and all equations need to be well specified.
- A reliable model may take years to build.
- Construction of these models requires data to be available at a very detailed level - it may require the bespoke analysis of raw data.

General equilibrium (GE) models

These are a form of structural model. They differ from standard structural models which are based on partial equilibrium assumptions in that they take into account interdependencies of markets to model general equilibrium in all markets. **Computable General Equilibrium (CGE) models** are large-scale, economy wide numerical simulation models. They account for economic interdependencies between different sectors of the economy and different agents through a set of complex equations. The structure of these models is highly flexible and allows for the evaluation of policy impact on different firms, households and the government finances.

Strengths

- Present "big pictures", i.e. directional impacts, and in capturing both direct and indirect policy effects.
- Can look at the sectoral impacts, tax incidence issues and distributional impacts amongst different household types.
- Key assumptions made by a CGE model can readily be evaluated through detailed sensitivity analysis.

Weaknesses

- Computational and conceptual complexity.
- Time consuming to build.

- Based on strong assumptions and require large amounts of data.

Cost-benefit analysis (CBA): quantifying in monetary terms all of the costs and benefits of a policy/programme. Includes quantifying wider social costs and benefits e.g. costs associated with air pollution such as health effects. Ideally CBA should be used in combination with all of the methodologies described, to enable one to make decisions and compare policy options in monetary terms.

Strengths

- Promotes consideration of the full range of policy effects by providing a useful format for bringing all of the costs and benefits together. Allows an assessment of the overall net worth of a policy.
- Enables different policies to be compared.

Weaknesses

- Not all social costs and benefits can be given monetary values. Some can be estimated through indirect means and converted into monetary values, but this will not be possible for all costs and benefits. Failure to estimate all significant costs and benefits adequately in monetary terms may distort results, giving rise to misleading conclusions and consequently the wrong decisions being taken.
- May not show whether a policy has achieved its objective (i.e. the policy may not be effective/successful in this context, even if it does have a positive benefit to cost ratio which exceeds that of the control group).

Natural experiment

Monitoring of information may contribute towards process and impact evaluation questions. It includes: i) analysis of administrative data and management information systems, ii) collection and analysis of performance measurement data, and iii) information form special monitoring exercises which have been put in place. Monitoring alone is not a form of evaluation, but provides essential information for conducting an evaluation.

Strengths

- Can be very cheap if the data is already available in current systems.
- Even if the original question is not fully answered it can give an indication of progress towards achieving an objective/highlight any major issues encountered.

Weaknesses

- Data in currently in systems may be recorded to a poor standard or be biased.

Problem structuring methods (PSMs) such as Oval Mapping Technique, Strategic Choice, Soft Systems Methodology (SSM), Strategic Options Development and Analysis (SODA) and the Delphi Technique. Typically Operational Researchers use these methods to help experts bring together their ideas in a structured way. In an **Oval mapping** workshop, ideas are written on ovals by the participants, then linked together. Sometimes the ideas are ranked by importance and used to work out an action plan. Sometimes other shapes are used instead of ovals, such as hexagons. For the **Delphi technique** the evaluator facilitates a group decision so experts can come to a consensus of opinion. It can be carried out by email or in a workshop and involves experts brainstorming ideas and then voting on the ideas to come to a relative order of importance. The results of the votes are fed back to the experts who are given the opportunity to change their votes. This is an iterative process and there can be several rounds of voting and feedback.

Strengths

- Used where there is no or little data available, and so are often known as "soft" techniques.
- Can combine expert opinion that is less biased by rank, seniority or “strong” personalities.

Weaknesses

- Not based on hard data so there needs to be follow up work to confirm or expand on the conclusions.

Quantitative surveys collect quantitative data. Usually large-scale and can have a role in addressing both impact and process evaluation questions. If surveys are required to collect baseline data they need to be commissioned before the policy is announced/ implemented. Surveys may be conducted by post, telephone or face-face.

Strengths

- Provide hard data on which to base the evaluation.
- Provide data that cannot be obtained through administrative or management systems.

Weaknesses

- Can be very expensive especially if they need to be repeated at different time periods.

In-depth interviews enable full exploration of personal experiences of interventions. Must be conducted by professional qualitative interviewers. The interviewer has a topic guide to ensure all the ground is covered and focus maintained. Flexibility is needed for follow-up questions which facilitate the in-depth exploration of accounts. They are interviewee led rather than interviewer led.

Strengths

- Rich resource of personal beliefs and experiences. Explicit explanations of participants’ views and reasons for their decisions and actions.
- May uncover effects of the intervention the evaluator had not thought of.

Weaknesses

- Interviewees may be unable or unwilling to give an open account of themselves, or find self questioning difficult.

Focus Groups/Group Discussions

These are discussions which bring together different recipients or participants to discuss a topic or set of topics. Typically consists of four to eight people facilitated by a researcher. A tape recorder often records the discussion.

Strengths

- Work well in combination with other research methods e.g. interviews.
- Hearing others’ views may improve the articulation of participants’ own views and experiences. Can help to highlight common experiences and views as well as differences.
- Provide opportunities for creative thinking to tackle abstract and conceptual topics or where an interview might not get as far.
- Can create an environment which stimulates solutions, strategies and ideas for improvement of an intervention.
- Can give immediate feedback on new ideas.

Weaknesses

- Not appropriate when the presence of others are likely to inhibit, constrain or influence others in ways that will make the accounts less open and reliable.
- Less useful when exploration of detailed personal accounts is required or of sensitive issues.
- Do not obtain such an in-depth picture as personal interviews.

Case studies

Case studies bring together multiple perspectives from key people. They differ from interviews and focus groups that give accounts of individuals seen from their perspective. Case studies provide a holistic and comprehensive in-depth understanding of experiences and outcomes. The case in question may be an area, client, office etc.

Strengths

- Powerful approach to understanding an intervention and the context around it.

Weaknesses

- Can be expensive, time-consuming and complex – although the smaller scale ones may not be more costly than other qualitative techniques.
- Often context-specific so difficult to generalise from.

Other process methods: There are several other research methods which can be used to carry out process evaluation. These include literature reviews and documentary analysis, consultative methods, biographical and life history approaches and the observation of a project by researchers.