# Using comparison group approaches to understand impact

## Contents

**Summary**

- Comparison group approaches involve comparing the outcomes of one group of service users with the outcomes of a different group, or groups, to give a better understanding of whether an intervention[1] has achieved its intended outcomes.
- Researchers rate comparison group approaches highly: they are seen as the best way to know whether an intervention has made a difference, and can help to attribute impact to an intervention. Because of this, many stakeholders, including government and other funders, want to see more charities conducting comparison group studies. However this should not be taken to mean that comparison group studies ought to be pursued at the expense of other approaches like theories of change or qualitative research. Rather they should be seen as adding to these by providing more concrete evidence of impact.
- A good comparison group is as similar as possible to the group of service users who are receiving an intervention, thus allowing you to be confident that the difference in outcomes between the groups is *only* caused by the intervention. Variables like age, gender and offending history can be used to check the similarity of a comparison group. However, the best way to ensure a fair comparison is to randomly assign people to one group or the other before the intervention is delivered. This means that all the factors that may affect results will be distributed equally between the two groups, with the only difference being the intervention itself.
- If random assignment is not possible there may be other ways to create comparison groups, but these methods will require high quality information about those receiving the intervention *and* those not receiving the intervention so that you can be sure the comparison is valid.
- Comparison group approaches are sometimes seen as difficult, costly or unethical, but these challenges can be overcome. This guidance describes different comparison group approaches and invites you to consider whether there are opportunities to conduct a study in your own organisation. However, we also stress that it's important to get expert help before going too far as there are a number of pitfalls that can lead to biased or flawed designs. You may be able to access free advice from universities or Government Departments.

[1] The word 'intervention' is throughout this guidance to mean any kind of programme, project, service or approach you are interested in evaluating: This could range from an advice session or arts-based intervention in prison to more comprehensive services such as supported housing.

## Introduction

This guide looks at research approaches that use comparison groups to measure and compare the difference made by voluntary, community and social enterprise organisations (VCSE). This includes research in which outcomes for groups of service users are compared to outcomes for groups that have not used a service, or have used different services. It also discusses research where a comparison groups are derived statistically, known as quasi-experimental approaches.

Good comparison group studies provide a high standard of evidence[2], but compared to other evaluation approaches they are harder to get right. It's therefore worth familiarising yourself with other approaches to evaluation such as theory of change, qualitative research and outcomes tools *before* considering comparison groups.

The aims of this guide are to:

- Introduce the concept of comparison groups and why they are seen as a high standard of research.
- Help you think about options you may have to use comparison groups to evaluate your own interventions.
- Highlight the issues to consider when thinking about comparison groups.

This guide does not provide detailed instructions on how to design and run studies, but does provide links to these if you are interested. If you do identify an opportunity to conduct a comparison group study for your work, we strongly recommend you seek expert advice as early as possible in the process to ensure the study is robust and useful.

## What are comparison groups?

A comparison group (also referred to as a control group[3]) is a group of people that does not receive an intervention, but in other ways is as similar as possible to a group that does. By studying this comparison group, you can estimate what would have happened without the intervention—often referred to as **the counterfactual**.

The underlying idea is that to really understand whether an intervention works, you must compare what happens when an intervention is in place and when it is not (or when a different intervention is in place). It is modelled on the way experiments are constructed in the natural sciences: Under laboratory conditions, it is possible to design experiments in which only one variable is changed, so that the effect of that variable can be measured. For example, you might have two flasks of water and apply heat to one (the 'treatment') and not to the other (the 'comparison'). Water in the heated flask boils, and because you know that in every other respect the flasks are the same you can conclude that heat *causes* water to boil.

[2] See our guidance on standards of evidence for a further discussion of this http://www.clinks.org/sites/default/files/StandardsofEvidenceGuide.pdf

[3] The two terms are not entirely interchangeable. A 'control group' tends to refer to groups defined through more robust approaches such as random allocation, while a comparison group is any group that can provide a reasonable comparison.

Although people and society are never as simple to control, the logic can still be applied. For example if you take two groups of offenders that are as similar as possible to one another, apply an intervention to one group (the 'treatment group') and not the other (the 'comparison group'), then measure the different outcomes of the two groups, you can then estimate the average effect of the intervention.

## Why are comparison group approaches so highly regarded?

Comparison group studies help to attribute results to interventions. To explain this, imagine that a charity working with a group of offenders finds that only 30% of its service users have reoffended. This may seem like a good result but in truth you can't tell whether it is or not, for two related reasons:

1. *You do not know what the reoffending rate would have been without the intervention.*

   You could look at the national reoffending rate but this would not be comparing like with like - reoffending rates vary for different groups. Alternatively you could use reoffending rates for more comparable groups or predicted rates (see discussion of OGRS scores[4] on page 11), but this still does not confirm that it is the intervention that is responsible for the difference because;

2. *Other factors may have influenced the result.*

   The user group may differ from the average by chance, some of them may have received another intervention, there may have been changes in police recording procedures, the local job market may have improved, or there may be other factors that have affected their chance of reoffending (e.g. it is widely known that as people get older they are less likely to offend). It's therefore impossible to tell conclusively that the intervention has made the difference and no amount of other information about the quality of the service, user satisfaction or stakeholders' views will be wholly sufficient to disprove alternative explanations, so claims about impact will remain tentative. This is referred to as the problem of **attribution**: even if you can see that two things are correlated—for example attendance at an intervention and reduced reoffending—you do not know that one causes the other; both may be caused by an underlying willingness to desist from crime.

Properly constructed comparison group studies attempt to resolve this problem by minimising the effect (or 'noise') of other variables. For example, if service users are assigned randomly to either the treatment or control groups, then external factors such as changes to the local job market will apply to both groups so can be ruled out as an explanation for difference in outcomes.

Consider a different intervention which works with offenders to improve self-esteem, for example through arts or volunteering. It is possible to measure self-esteem[5] at the beginning

[4] http://eprints.lancs.ac.uk/49988/1/ogrs3.pdf

[5] Please see our guidance on outcomes tools for more information on how to do this http://www.clinks.org/sites/default/files/UsingOffShelfToolstoMeasureChange.pdf

and end of the intervention to see if it improved (known as a pre & post design). A positive change is an encouraging result, but the problems highlighted above persist; you do not know how self-esteem would have changed without the intervention, or whether other factors have made a difference. Measuring the level of change in a well matched comparison group who do not receive the intervention would resolve this.

For these reasons comparison group approaches are often seen as the gold standard of research.[6] Many funders, commissioners and researchers regard positive findings from robust comparison group studies as the best proof that an intervention has made a difference. If studies are repeated and more evidence is found in different settings it is taken to show that an intervention "works", which can lead to more funding and wider roll-out. The chart below illustrates this by showing the widely-used NESTA standards of evidence. As you move through the scale, the credibility of different research methods increases, and a robust comparison group study is minimum requirement for 'level 3' evidence and above[7].

Nesta Standards of Evidence

| Level | Our expectation | How the evidence can be generated |
|---|---|---|
| At Level 1 | You can give an account of impact. By this we mean providing a logical reason, or set of reasons, for why your intervention could have an impact and why that would be an improvement on the current situation. | You should be able to do this yourself, and draw upon existing data and research from other sources. |
| At Level 2 | You are gathering data that shows some change amongst those receiving or using your intervention. | At this stage, data can begin to show effect but it will not evidence direct causality. You could consider such methods as: pre and post-survey evaluation; cohort/panel study, regular interval surveying. |
| At Level 3 | You can demonstrate that your intervention is causing the impact, by showing less impact amongst those who don't receive the product/service. | We will consider robust methods using a control group (or another well justified method) that begin to isolate the impact of the product/service. Random selection of participants strengthens your evidence at this Level, you need to have a sufficiently large sample at hand (scale is important in this case). |
| At Level 4 | You are able to explain why and how your intervention is having the impact you have observed and evidenced so far. An independent evaluation validates the impact. In addition, the intervention can deliver impact at a reasonable cost, suggesting that it could be replicated and purchased in multiple locations. | At this stage, we are looking for a robust independent evaluation that investigates and validates the nature of the impact. This might include endorsement via commercial standards, industry Kitemarks etc. You will need documented standardisation of delivery and processes. You will need data on costs of production and acceptable price points for your (potential) customers. |
| At Level 5 | You can show that your intervention could be operated up by someone else, somewhere else and scaled up, whilst continuing to have positive and direct impact on the outcome, and whilst remaining a financially viable proposition. | We expect to see use of methods like multiple replication evaluations; future scenario analysis; fidelity evaluation. |

[6] Please see our guidance on standards of evidence for a further discussion of this: http://www.clinks.org/sites/default/files/StandardsofEvidenceGuide.pdf

[7] http://www.nesta.org.uk/publications/standards-evidence-impact-investing

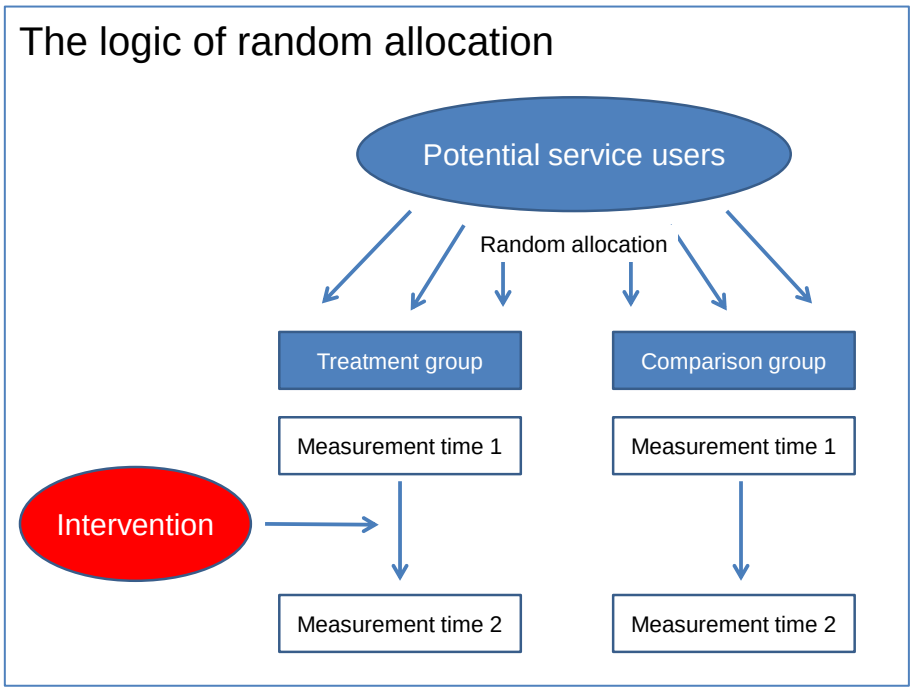
Although they are well regarded it's important to stress that comparison group studies do not supersede other forms of evidence or research, rather they add to and support them. Other forms of research are valuable because they give an earlier indication of success and more information about how things work. Moreover we always recommend that any evaluation, including comparison group studies, are underpinned by a clear articulation of the theory behind the intervention using approaches such as logic modelling or theory of change[8].

## Approaches to creating comparison groups through randomisation

Having discussed the principle behind the use of comparison groups we now describe the different methods that can be used to create them.

*Randomised Controlled Trial*

The Randomised Controlled Trial (RCT) is the most powerful approach to creating a comparison group. In its pure form it involves recruiting a group of potential service users and then randomly selecting some to receive an intervention and some not. Assuming there is a sufficient number in each group and that it is done correctly, the process of random assignment eliminates the possibility of external factors affecting the results, so that any differences between the two groups should be solely the result of differences in the interventions they did or did not receive. The logic of the approach is summarised in the chart below.[9]




[8] See our guidance on the theory of change approach http://www.clinks.org/sites/default/files/TheoryofChangeGuide.pdf

[9] Take from HM Treasury (2012), Quality in Impact Evaluation https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/190984/Magenta_Book_quality_in_policy_impact_evaluation__QPIE_.pdf

An example of RCTs in action is in clinical trials of new drugs. Two groups of patients with the same condition are randomly assigned, with one receiving a new drug and the other a placebo, which contains nothing useful.[10] Any difference in outcomes between the two groups should then represent the effect of the drug.

**Praxis Community Projects** (case study A) is currently conducting an RCT to test the impact of a mentoring programme for ex-offenders from Central and Eastern Europe. The details of eligible service users are sent to the Centre for Economic and Social Inclusion (CESI), and are randomly assigned to a 'treatment' or 'control group'. Over the coming months, the treatment group will be compared to the control group for differences in outcomes such as employment, training and reoffending, as well as attitudes and outlook.

In practice, randomisation can be hard to set up. Service managers may have little control over who is referred to an intervention. They may not have enough potential service users to deny the intervention to some of them, or stakeholders - such as funders, users and staff - may feel it is unacceptable for some people to be denied an intervention (see the discussion of ethical issues on page 15). However, in these situations there may still be other options.

*Varying the intervention*

Randomisation can be applied to variations in an intervention rather than the intervention as a whole. For example, one group could be assigned to receive weekly mentoring appointments and another to receive them monthly, or the provision of a limited number of voluntary work placements within a wider mentoring scheme could be allocated randomly. Similarly, if you want to change an element of your intervention you could apply the change with a randomly selected group of users while the rest continue to receive a normal service – telling you if the change has any impact.

Private companies often use this kind of approach, for example randomly assigning customers to different website designs to see which generates the most sales. Clearly, this is less useful at demonstrating the impact of an entire intervention, but it can still provide useful findings about the impact of particular elements or features.

*Waiting list designs*

Waiting list designs are a way to overcome ethical concerns associated with denying people services. In these, the intervention is delivered in phases, all service users get the intervention in the end, but random allocation decides who gets it now and who gets it later. The 'later' group therefore acts as a comparison group to the first cohort of service users.

This design is particularly useful when constraints on staff or resources mean that not everyone can access an intervention at the same time anyway. However, its weakness is

[10] A placebo is used because even the process of receiving and taking a drug may have an impact regardless of what's in it ('the placebo effect').

that you cannot measure the impact of an intervention on longer-term outcomes such as reoffending (as everyone will receive the intervention at some point).

**The Prison Phoenix Trust** (case study B) is an example of a waiting list approach. The study aimed to test the impact of yoga courses on prisoners' well-being and self-control. Researchers recruited a cohort of 167 volunteers from seven prisons, who were then randomly assigned to either a yoga group or a comparison group. Prisoners in both groups were assessed at the beginning and end of the first course and differences measured. The course was then delivered to those in the comparison group (for more information about this research see appendix 24).

*Borderline randomisation*

In this approach assignment to an intervention is done according to an assessment of need. For example, those with the highest needs will definitely receive the intervention, while those with the least needs definitely do not. Then there might be a third group, for which it is unclear whether the intervention is the best use of resources or not. People in this borderline group can be allocated at random to receive the service or not and their results can compared. Alternatively, if needs are assessed on a quantitative scale with a cut-off point for receiving an intervention, you could compare those around the cut-off point, who can be presumed to be relatively closely matched.

*Natural experiments*

Finally there may be times occasions where divergences in law, policy or practice offer the opportunity to analyse populations *as if they have been part of an experiment* – which are called 'natural experiments'[11]. For example, it may be that people released from the same prison receive different 'through the gate' interventions depending on which communities they are released back into. The results of the two groups can be compared, especially if analysis is conducted to check how well the two groups match (e.g. by variables such as age and offending history).

However, natural experiments are seen as less powerful than purely randomised approaches because of the risk of possible variations in other variables like the type of communities, offenders, and institutions such as police and courts - all of which could influence the outcomes of different groups and bias the comparison.

[11] For a good introduction see http://www.scotland.gov.uk/Resource/Doc/175356/0091395.pdf

## Quasi-random approaches for creating comparison groups

If randomisation is genuinely not possible, an alternative is to identify comparison groups from existing data or situations, using what are known as quasi-random approaches.

*Difference-in-difference designs*

In difference-in-difference designs, changes amongst service users are compared to changes in comparable groups that are *theoretically* the same. Therefore a precondition of the approach is that you have 'before' and 'after' data (for example around self-esteem or pro-social attitudes),for both your service users and another comparable group.

Cohort-based studies are an example: if an intervention is new, so earlier cohorts did not receive anything similar, then change in service users can be compared to that found in earlier cohorts. Similarly, you could compare your users to populations in other areas. For this to work well there must be little or no other changes or differences in context that could affect the outcomes. For example if you run an intervention across a prison and then compare reoffending rates to those achieved before the intervention you will need to be sure nothing else has changed (such as a change in the type of prisoners held there).

A similar type of difference-in-difference design is to compare rates of change in the *same* group. For example, if regular data was being collected on a prison group (e.g. Oasys scores or negative behaviour entries) so that you had a least three historical data points, you would know what the existing trend in that data was. You would then apply the intervention while continuing to measure the same data and assess whether the trend seems to be altered. Hence, the comparison group is the same group, only before you worked with them.

Approaches like this clearly depend on the availability of good data on so that you can compare rates of change. Even with this in place, you may not be able to control for variables in the external environment which ultimately means that uncertainty about the findings will always be more pronounced than for a genuinely random study.

*Drawing theoretical comparison groups from available datasets*

A different type of quasi-random approaches is those that draw treatment and comparison groups from existing databases. An example of this is the Justice Data Lab[12] which takes organisations' own records of who they have worked with to identify the treatment group from within Police National Computer (PNC) dataset, and then uses a statistical process called 'Propensity Score Matching' to define a comparison group from the same database which matches the characteristics of the treatment group as closely as possible. Helpfully, there is no limit to the size of the comparison group that can be defined, so a relatively small cohort of 30 service users can be compared to many thousands of similar people, which improves the accuracy of comparisons.

[12] For more information on this please go to http://www.clinks.org/sites/default/files/MoJ%20Data%20Lab%20briefing.pdf

However, the comparison group is only matched on a few variables so there may still be important differences between it and the treatment group. For example, there is no variable on the PNC for substance misuse, so if an organisation focuses its work on drug users the Justice Data Lab cannot define an accurate comparison group.

**Prisoners Education Trust (PET)** provides grants to offenders for distance learning courses or to purchase materials for arts and hobbies. They sent details of all prisoners to whom they had provided grants between 2002 and 2011 to the Justice Data Lab, which was able to identify a 'treatment' group of 3,091 offenders from the PNC. The reoffending rate for this group was 19%, compared to 26% for a control group matched by key characteristics. It means PET's intervention is associated with an estimated reduction in reoffending of 7 percentage points (or around a quarter).[13]

*Regression modelling*

Another statistical approach is regression modelling. As above, this depends on having a good enough and large enough service user database, including a 'dependent variable' (e.g. reoffending) and a range of 'independent variables' (e.g. level of participation in initiatives, criminal history age and gender).

The regression process then tests associations between the independent and dependent variables, while controlling for other variables. For example, you might find that an intervention initially appears to be associated with a reduction in reoffending, but when controlling for age this effect is reduced because uptake of the initiative was higher amongst older offenders who were less likely to re-offend anyway.

The weakness of regressions is that you can never control for *all* the variables that may have influenced an outcome, whereas in a RCT design you can assume that even unknown variables have been equally distributed across comparison groups[14]. Nonetheless, if you have, or can create, a good enough database, then regression is well worth trying, although you may need someone with knowledge of maths or statistics to help.

## Other approaches

Apart from random and quasi-random approaches, there are other ways to define comparison groups that are less powerful but may still provide useful information.

*Comparing outputs and outcomes across settings*

Comparing outputs and outcomes from natural variations in the services you provide can help you to better understand what does and does not work for different people. For example, you might see different outcomes across a number of prisons you work in and

[13] https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/270084/prisoners-education-trust-report.pdf

[14] It is worth noting that regression has a range of applications and is integral part of the analysis of randomised control trials.

study the reasons for this. These are not comparison groups as such, but the process of studying differences will provide you with an increased understanding of different 'outcomes patterns' that may help you identify the key ingredients that make your intervention work and help you improve or refine the services you provide.

*Comparing service users across partners*

You may be able to talk to your referral partners about which service users they send to you and which they send to other providers; to see if their outcomes can be compared. However, you would need to be cautious when making these comparisons as it is unlikely your partners will have referred people on a random basis—they will most likely have considered who will benefit from which intervention, which means your service users are already different by their propensity to benefit from what you do. Of course, if you can persuade your partners to assign service users randomly to different providers then this is the basis for an RCT and for much more powerful analysis and understanding

If you do compare the outcomes for your service users with those of other providers the quality of the analysis can be improved by looking to see how well matched the groups are by characteristics like age and gender and offending history - which offers a limited way to ensure similarity between comparison groups. Moreover, you may find options to conduct more targeted and robust comparisons; for example identifying a subgroup of your users who are a better match to users of another service, or you could specifically match individuals who have used your service to individuals within a referral partner's wider caseload (as in the example below).

**The Brathay Trust** (case study D) saw an opportunity to improve its understanding of impact by developing comparison groups from amongst young people that were *not* referred to them by a Youth Offending Team (YOT). Each young person referred to Brathay was matched to another young person by the YOT according to key characteristics (a process called 'stratified matching'). Analysis was then conducted to compare the two groups by outcomes in criminal conviction and school exclusions, to see whether referral to Brathay was associated with any differences.

*Offending Group Reconviction Score*

Charities sometimes make the error of comparing their reoffending rate outcomes to published national reoffending rates or those of particular regions. However, as noted above, offending rates differ considerably from group to group, so comparisons like this can be misleading. In practice the likely reoffending rate of any group of offenders is unique, which is why comparisons should really only be made to comparison groups that have been defined by more robust approaches.

Using the Offending Group Reconviction Score (OGRS) is probably the easiest way for a charity to make these more robust comparisons This is a tool widely used in probation which estimates the probability that offenders *with a given history of offending* will be sanctioned for any new recordable offence within two years of noncustodial (including suspended) sentence

or release from custody [15]. In OGRS, the likelihood of reoffending for specific cohorts or individuals is calculated from historical analysis of the rates associated with static risk factors (such as age, gender and criminal history) - providing a figure that can be compared to the actual reoffending rate for a group of service users.

The potential weakness is that there will still be unknown differences between your service users and the OGRS average. Despite this it often represents the best way for organisations to compare their outcomes to official statistics and is regarded by NOMS as a "reliable and valid predictor, suitable for use in commissioning analysis" (although the Justice Data Lab is technically superior because of the direct matching of user and comparison groups).

**TheHorseCourse** is an equine-assisted behaviour programme delivered in prisons and other settings, including working with 25 prisoners at HM Portland. The programme manager has accessed Prisoner National Offender Management Information System (P-NOMIS) data to determine that the one-year reoffending rate of their service user group was 44%, compared to an OGRS predicted rate for the group of 63%; an indicative fall of 19 percentage points. Because the sample is small, TheHorseCourse cannot say with certainty that this reduction in reoffending is significant, but taken alongside other evidence collected, the programme stands out as a promising intervention with apparent success among some of the hardest to engage offenders.

*Comparing those who engage with the service and those who do not*

Another way to construct a comparison group is to look at differences between people that volunteer and engage with your intervention and those that do not. The is the least valid approach because the factors that made some people participate (like willingness to change and higher aspirations) are also likely to have a strong effect on the outcomes you are aiming to achieve, (like reduced reoffending), so the and treatment groups are not actually the same.

*Self-reported counterfactual*

The final way to achieve a 'theoretical counterfactual' is to simply ask service users what would have happened without the intervention. This is always worth doing because it can provide some illuminating insights and case studies about how an intervention works but it's important to remember that the findings will not be seen as particularly strong. This is because service users do not really know what would have happened otherwise, they can only guess, and there is also a strong tendency for them to overstate the effect an intervention because they think it is want you want to hear (this is known as 'social desirability bias').

[15] http://eprints.lancs.ac.uk/49988/1/ogrs3.pdf

## Limitations and criticisms of comparison group approaches

Despite being regarded as the gold standard of research there are some limitations or issues associated with comparison group approaches.

*Understanding **why** and **how** an intervention works.*

Often the results from comparison group studies are one-dimensional; a single 'net effect' of the intervention. This helps to provide proof of concept, but other research methods are needed to understand what has produced this result, especially when dealing with complex questions such as what encourages people to desist from crime. A related criticism is that RCTs are often focused on specific programmes—with a start and end date—such as mentoring schemes or courses. However, these programmes include a whole range of practices and activities whose effectiveness are less easily measured; like staff skills; the sequencing of support; or the quality of relationships established with offenders. Some people argue that by focusing too much on accrediting or proving structured "programmes", we pay too little attention to the complex and dynamic on-going processes that actually make these programmes work.[16] This is why many consider the real 'gold standard' of evaluation to be an RCT ('does it work?') combined with in-depth qualitative research ('*how* and *why* does it work?').

*Applying the findings to different services and contexts*

Comparison group approaches only test whether something has worked in a particular context, and with all the specific practices and conditions of that intervention in place. But society is always changing, and even within a particular context, an intervention may work for some people but not others. You therefore need to consider whether findings can be generalised to different settings (an issue known as 'external validity').[17] On the other hand, advocates of comparison group approaches argue that these issues can be overcome if studies are repeated to provide greater confidence. This process of accumulation and synthesis is known as 'systematic review', whereby researchers study the effects of interventions across different settings to draw stronger conclusions.[18]

*Programme variability*

Another drawback is that for complex programmes, measuring success (or repeating any success found through comparison group approaches) requires tight control over programme

[16] This perspective is discussed in Hough, Mike (2010) *Gold standard or fool's gold: the pursuit of certainty in experimental criminology* http://eprints.bbk.ac.uk/3815/1/3815.pdf

[17] See, for example, Pawson (2013) The Science of Evaluation: A Realist Manifesto. Or other commentaries:

http://www.europeanevaluation.org/images/file/Member_section/EES%20Statement%20on%20Methodological%20Diversity/EES%20Statement.pdf

http://thecomingprosperity.blogspot.co.uk/2011/09/why-randomized-controlled-trials-work.html

http://www.ssireview.org/blog/entry/rcts_not_all_that_glitters_is_gold

[18] The Campbell Collaboration in the US are leaders in the areas of systematic review in the social sciences. For example, they have e conducted a review of mentoring projects in criminal justice
http://www.campbellcollaboration.org/news_/Promising_results_mentoring_programs.php

delivery. Variations on the can ground mean you are not measuring the same things across different clients so you will not learn how to repeat previous success. The need to control delivery can lead to additional instructions, training and bureaucracy which may increase costs while stifling the flexibility and professionalism of frontline staff.

*Resources*

Finally, it is often argued that a robust comparison group studies requires resources and expertise beyond most charities. Whilst this is undoubtedly true, there are opportunities , for organisations to conduct them relatively easily by being creative and manipulating existing systems and databases to find comparison groups (as illustrated by the case studies in this document). Furthermore, funders are increasingly interested in resourcing these kinds of studies as they will help demonstrate their own impact and guide future funding decisions.

## Ethical concerns

Comparison groups—and in particular RCTs—can raise ethical concerns because they involve deliberately denying people an intervention that is intended to help them. However, a strong counterargument is that if you are uncertain whether an intervention works, or even whether it might make things worse, it is *more* unethical to continue resourcing and spending money on it until it is properly tested. Therefore RCTs should be seen as ethical unless the benefits of an intervention are already known, or when the impact of refusing a service will have a detrimental effect on a person's wellbeing irrespective of whether it is worthwhile (for example if a prisoner is denied something they wanted this may affect their wellbeing or morale[19]).

A good illustration of why RCTs are ethical is to consider "**scared straight" programmes**. These involved organising visits to prisons by young offenders, and were aimed at deterring them from criminal activity. A number of comparison group evaluations were conducted and then reviewed by the Campbell Collaboration. The authors found that these programmes were actually more likely to have a harmful effect and increase risk of offending, relative to doing nothing at all. It would have been unethical to continue these programmes without knowing their effect.[20]

A separate ethical issue is what you tell participants about the process. Whichever type of research you are conducting, participants should always give informed consent to take part in a study. For comparison group studies, it is important to explain that an evaluation of a new method is taking place—not an evaluation of a better method, as if you knew it was better, you would be doing it already.

However there are also occasions when it is not necessary or helpful to inform people they are part of a study, for example if you are testing different forms of communications materials. In short, there are no fixed ethical rules to apply but we encourage you to think

[19] Which would also make them a poor comparison and therefore, arguably, lead to a flawed study

[20] See the following for more information http://www.campbellcollaboration.org/lib/download/13/

through the implications of any research you are conducting and be open and consult widely if you have concerns.[21]

## Conducting a randomised control study – the process and issues to consider

In the preceding sections we have highlighted some of the different ways a comparison group study might be constructed, ranging from the most robust (random allocation) to the least (self-reporting and comparing those who engage to. those who don't). The aim has been to help you think about any opportunities you might have to conduct a comparison group study, as well as interpret the quality of any comparison group studies you read about.

If you think you have opportunities to evaluate your services with a comparison group approach, we recommend that your next step should be to get some expert help—for example from a funder, university criminology department, research organisation or a Government department (e.g. Ministry of Justice or the Cabinet Office Behavioural Insights Team).[22] This is because negotiating the finer points and pitfalls of setting up a good comparison group study usually requires expertise and, importantly, time to ensure that all the arrangements for sharing information are in place.

**The Prison Fellowship** has commissioned a Randomised Control Trial of the Sycamore Tree programme in which Prison Fellowship volunteers teach the principles of restorative justice to groups of up to 20 learners in prisons. To set up the study they worked with a PhD student from the Jerry Lee Centre of Experimental Criminology at the University of Cambridge. This has meant that, at a relatively low cost, they have access to high-profile academics and are able to conduct a very high quality study.

[21] For further discussion of ethical issues please see section 7 of our guidance on engaging service users http://www.clinks.org/sites/default/files/AcheivingUserParticpationResearch.pdf

[22] https://www.gov.uk/government/organisations/behavioural-insights-team

*Nine step process for conducting RCTs*

While we recommend expert help, to give you a sense of what is needed to run a study, we recommend looking at the 2012 Cabinet Office publication *Test, Learn and Adapt*,[23] which contains an enthusiastic endorsement of RCTs and a nine-step process for conducting them. These steps are briefly described below with reference to relevant examples from criminal justice:

| **Nine steps for running a RCT, adapted from Test Learn and Adapt (Cabinet Office, 2012)** | |
|---|---|
| ***Testing stage*** | ***Issues to consider*** |
| **1. Identify two or more policy interventions to compare.** | Here 'interventions' may refer to a whole programme or to relatively minor changes to existing programmes. 'No support' or 'business as usual' can also be regarded as an intervention to test[24].<br><br>Before embarking on a comparison group approach you must review the available literature; be up-to-date on what is already known in the area; and pilot the intervention to test whether it seems to work and iron out any problems[25].<br><br>Ideally, comparison group studies are best done at the point when an intervention seems ready for testing but *before* it is rolled out more widely. |

[23] https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/62529/TLA-1906126.pdf

[24] Although. 'No support' is probably very unlikely in criminal justice, so you are more likely to be comparing with 'business as usual', i.e. the package of interventions people in the system typically receive.

[25] Our web page links to sites through which you can start to engage with existing literature http://www.clinks.org/support-evaluation-and-effectiveness/existing-evidence.

| | |
|---|---|
| **2. Determine the outcomes and how they will be measured.** | Any good comparison group study is based on an understanding of outcomes and how these will be measured. The most widely used outcomes data is official data such as reoffending rates, but self-reported offending rates have also been shown as robust.[26]<br><br>There are also a number of intermediate outcomes that have an established relationship with reducing re-offending,[27] for example in the Phoenix Prison Trust case study a range of cognitive and well-being outcomes were used. To determine intermediate outcomes we recommend developing a theory of change to identify and agree on the most important outcomes for your service.[28]<br><br>You will also need a robust process for measuring outcomes. This may be a way to measure the reoffending rate for service users, or, if you are focusing on intermediate outcomes, tools that are used 'before and after' the intervention[29].<br><br>The outcomes you choose to measure should be plausibly connected to the scale of the intervention. For example, if you are running a small-scale arts-based initiative you should probably not choose reoffending as your outcome because it is too remote and dependent upon on other factors. Better outcomes might be intermediate cognitive, attitudinal or behavioural changes that have been shown by other research to be linked to desistance from crime.[30] |
| **3. Decide on the randomisation unit.** | Different units can be randomised. Often it is individuals who are assigned to one intervention or another, but the process can equally be applied to groups of users, institutions or regions.<br><br>For example, an intervention may be operating in a number of prisons, but within a prison there may be no possibility to select which prisoners will or will not take part. In this case you might decide to deliver the intervention in some prisons, or in some wings, and not in others on a random basis and compare differences in the aggregate outcomes across the two groups. |
| **4. Determine how many units are required.** | Comparison group studies are only valid if there are sufficient numbers in both the treatment and comparison groups for robust comparison. For example, if there are only 10 people in each then differences in outcomes are as likely to be the result of chance as the effect of the |

[26] See, for example, Farrington 2001"What Has Been Learned from Self-Reports about Criminal Careers and the Causes of Offending?" http://www.crim.cam.ac.uk/people/academic_research/david_farrington/srdrep.pdf

[27] These need to be supported by reference to existing academic research. See http://www.clinks.org/support-evaluation-and-effectiveness/existing-evidence.

[28] http://www.clinks.org/sites/default/files/TheoryofChangeGuide.pdf

[29] Please see our guidance on 'off the shelf tools' to measure intermediate outcomes associated with reoffending http://www.clinks.org/sites/default/files/UsingOffShelfToolstoMeasureChange.pdf

[30] A good review of this evidence is https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/243718/evidence-reduce-reoffending.pdf

| | |
|---|---|
| | intervention.<br><br>Statistical power calculations[31] should be used to calculate the size of sample needed to determine whether any difference is a result of chance or the intervention. These are dependent on the possible size of the effect (or, the difference or 'signal' you expect) and the size of the sample(s). So if an intervention is only expected to have a small effect on outcomes, a larger sample size will be needed to detect it with any degree of certainty. |
| **5. Assign each unit using 'robust' methods.** | As discussed above, the most robust way to assign a control group is to assign participants to different interventions randomly. This is best done by someone who is not working on the frontline as they may bring judgement and bias into the selection process. If you have a database of possible service users then randomisation can be done using the RAND function in Excel, or alternatively http://www.randomizer.org is an easy-to-use tool for assigning participants.<br><br>In non-random designs you need to think carefully about selection bias—the possibility that treatment and comparison groups are different in a way that might affect the results[32]. Therefore groups need to be compared by as many variables as possible to ensure equivalence[33]. |
| **6. Introduce the interventions.** | This stage refers to the delivery of the intervention itself. If it is delivered across a number of sites or by a number of different staff or volunteers it will be important to ensure as much consistency as possible, both so that you can better understand and describe the intervention itself and so that possible effects are constant.<br><br>If people drop out of the programme their outcomes should still be included and analysed, this is because drop-outs represent the attrition that the programme could expect in future and therefore will provide a more realistic estimate of future impact.<br><br>Similarly, assuming your aim is to test the impact of an intervention to make a case for future rollout it will be important that the intervention is as close as possible to that which will be eventually delivered at scale.[34] |

[31] http://wise.cgu.edu/powermod/

[32] You can read more about this issue in our guidance on sampling http://www.clinks.org/sites/default/files/IntroductionToSampling_0.pdf

[33] In fact this should always be done, even in random designs, as a safety measure.

[34] A common problem is that interventions often perform better at pilot or testing stages, possibly because the staff involved are more enthusiastic at earlier stages.

| *Learning stage* | *Issues to consider* |
|---|---|
| **7. Measure the results.** | At this stage, results are assessed to estimate the difference an intervention has made. This calculation is done statistically and there will always be margins of error around any estimate, particularly if sample sizes are small.[35]<br><br>It is also necessary to consider how long desired outcomes might take to appear before they can be measured. The Ministry of Justice Data Lab uses a one-year reoffending rate, and these typically take a minimum of 18 months from release from custody (or the point that a community intervention starts) for this data to become available. |
| *Adapting stage* | *Issues to consider* |
| **8. Adapt your policy.** | The ultimate aim of all comparison group design studies is to test what works so that services can be continuously improved. It is therefore important to act on the findings; in particular, if a study has shown limited evidence for impact then it is important to understand why, and to make changes, or possibly discontinue it altogether.<br><br>If you intend to report or publish your findings, it is important to be transparent and accurate in your description of the methodology.[36] It is every bit as important to report 'no effect' or negative findings as it is to report positive findings – as this is how everyone learns and avoids repeating ineffective interventions or practices. |
| **9. Go back to the beginning and continuously improve.** | The final stage reflects the point that It is good practice across all evaluation to continue to "test and learn". For comparison group studies there is the broader aim of accumulating and refining our collective knowledge of 'what works'.<br><br>A good example of this in action is Cognitive Behavioural Therapy (CBT) where many years of repeated RCTs across a variety of settings and service types have helped professional psychologists learn and improve. As a result, CBT is largely considered *proven*, and it's now the main approach commissioned in the NHS for a variety of disorders. |

[35] For example, the margins of error reported in Justice Data Lab findings tend to be quite high because of the relatively small sizes of the treatment groups (see some of the reports shown here https://www.gov.uk/government/collections/justice-data-lab-pilot-statistics)

[36] We recommend seeking expert help to do this. You can also refer to our guidance on writing evaluation reports http://www.clinks.org/sites/default/files/ReportWritingGuide.pdf

## Conclusion

Most evaluation conducted by or on behalf of charities is still at the lowest rung of the standards of evidence. This is partly due to issues like weaknesses in sampling and questionnaire design, but the chief problem is the lack of studies with a comparison group or counterfactual, which makes providing evidence of effectiveness extremely difficult. Of course a key reason for this is a lack of funding for evaluation, but even with this constraint there may still be opportunities to conduct comparison group studies and identify counterfactuals along the lines we have highlighted in this guide.

There is a lot of debate about comparison group approaches and RCTs in particular, but many of these issues can be side-stepped if they are regarded as a complimentary tool alongside other approaches. In some circumstances they are potentially very powerful, but this should not stop you from pursuing other forms of understanding and evaluation such as theories of change and qualitative research.

Conducting comparison group studies can be complex and are difficult to undertake without expert advice. This is important to remember because a badly-designed comparison group study is worse than none at all (because it costs money without showing anything useful). Indeed, you may even find, as some charities have, that engaging with universities and other organisations may help you to access funds and resources to support your aims.

## Further resources

An accessible introduction and manifesto for control group approaches is the Cabinet Office Behavioural Insights Team publication, *Test Learn and Adapt*: https://www.gov.uk/government/publications/test-learn-adapt-developing-public-policy-with-randomised-controlled-trials

The Magenta book provides more detail on the government's perspective and approach to impact evaluation, which is firmly based on the value of comparison group approaches (and random allocation in particular). It also contains a wealth of technical detail: https://www.gov.uk/government/publications/the-magenta-book

Another accessible introduction, written for the schools sector, is the Education Endowment Foundation's *DIY Evaluation Guide*: http://educationendowmentfoundation.org.uk/library/diy-evaluation-guide

Better Evaluation has a number of useful and accessible pages on the topic: http://betterevaluation.org/plan/understandcauses/compare_results_to_counterfactual

An article introducing the benefits of RCT designs: http://www.investinginchildren.eu/blog/rct-or-not-rct

This is accessibly written report, reflecting the perspective of a charity or housing association on the front line. Section 3.2 is most relevant: http://www.midlandheart.org.uk/displayfile.asp?id=57557

An article from the US, introducing and advocating for RCTs and highlighting some relatively low cost examples: http://coalition4evidence.org/wp-content/uploads/Rigorous-Program-Evaluations-on-a-Budget-March-2012.pdf

An article examining the design of evaluations in settings where there is a choice as to how an intervention is to be introduced and evaluated: http://eprints.port.ac.uk/9360/1/Parable_01.06.09_Last_submitted.pdf

A presentation discussing some of the more technical approaches that can be used to construct comparison groups: http://www.ifs.org.uk/docs/Dearden_SRA_December%206%202011.pdf

Finally, an article and presentation outlining an alternative perspective, and the limitations of comparison group approaches:

http://eprints.bbk.ac.uk/3815/1/3815.pdf

http://www.worldcongressonprobation.org/uploaded_files/Plenary-Desistance-Research-and-Evidence-based-probation-Farrall-McNeill-Maruna.pdf

## Appendices: Case studies

**Case study A: Praxis Community Projects** is conducting a randomised comparison design study to test the impact of a mentoring programme for ex-offenders from central and eastern Europe.

Potential users of the service are drawn from the caseload of the London Probation Trust. Service users who are eligible are invited to take part, and, if they accept, their details are sent to the Centre for Economic and Social Inclusion (CESI) who randomly assign people to either a treatment or comparison group. Those in the comparison group are informed and asked whether they will still be willing to be part of the evaluation—so far most have accepted.

Over the coming months, the treatment group that is receiving the mentoring programme will be compared with the comparison group for differences in hard outcomes such as employment, training and reoffending, as well as soft outcomes such as attitudes and outlook, using a tool developed by CESI.

A key challenge for the researchers is to recruit enough people to both the treatment and comparison groups to enable robust comparisons. A target of 100 people per group is the aim, which should be sufficient to determine whether the mentoring programme is associated with statistical difference in outcomes.

Alongside the random comparison element described above, the programme is being studied using other methodologies to examine the views and experiences of staff delivering the service and the perceptions of service users.

**Case Study B: Prison PHOENIX Trust—Measuring the impact of Yoga and meditation in prisons**

The Prison Phoenix Trust (PPT) encourages prisoners in their spiritual lives through meditation and yoga. It offers individual support to prisoners and prison staff through teaching, correspondence, books and newsletters. PPT works in prisons, young offender institutions, immigration removal centres, secure hospitals and probation hostels.

The potential benefits of yoga for prisoners include reductions in anxiety and aggression, and improved cognitive processes, like memory, attention and self-comparison—all factors seen as linked to desistance from crime.[37] Previous studies in neuroscience and psychology have provided some evidence for these links, but none have systematically measured the outcomes achieved by yoga courses amongst prison populations.

This gap in the evidence base prompted PPT to think how it might collect robust data to test its belief that yoga is valuable—and, depending on the results, persuade more prisons to invest in it. They spoke to academics at Oxford University who developed the research design and secured additional funding. The research was a collaborative effort; with PPT setting up yoga classes in prisons and arranging meetings between prisons and researchers, and Oxford University conducting the research fieldwork, analysis and write-up of the findings.

The researchers recruited a cohort of 155 volunteers from seven prisons, who were then randomly assigned to either a yoga group or a comparison group. Those in the comparison group were told they would be offered a place at a later date, which helped mitigate ethical concerns. The profile of the two groups was compared to ensure they were matched by age, gender and other variables (which is an important and necessary part of any RCT).

Yoga classes were run by trained teachers, once a week for two hours. Participants in both groups were asked to keep exercise diaries, which for the yoga group included time spent practicing yoga outside of class.

A number of psychological assessments were made of both groups before and after the intervention. At the end of the study all participants were given a cognitive behavioural task to assess attentional capacity and behavioural response inhibition. The results showed

[37] See http://www.justice.gov.uk/downloads/about/noms/commissioning-intentions-2013-14-oct12.pdf page 11

positive and statistically significant differences between the yoga and comparison groups, providing the first robust evidence for the benefits of yoga in a prison population.[38]

Conducting the research was not without challenges. From the initial idea, securing funding, arranging access to prisons, gaining ethical approval and writing up the results, the process took between three and four years. There was also an important methodological challenge to minimise attrition from both the yoga and comparison groups.[39] Now that the research is published, however, we have a better understanding of what yoga can achieve, while the Prison Phoenix Trust can be more confident about its results when talking to funders.

**Case Study C: Understanding the Impact of Care Farms**

Care Farms (CFs) use farming activities as therapy to help build self-esteem, improve physical and mental health, develop skills for employment and increase ability to interact socially. However, research into this approach is limited, with no large trials or other well-designed quantitative studies being conducted.

To address this, a pilot study has been designed by the University of Leeds to find the best ways to assess whether CFs can achieve their intended outcomes. 150 offenders serving Community Orders in a CF will be recruited, alongside a comparison group of 150 offenders serving them in other settings (for example building work, food handling, painting and decorating, recycling and cleaning). The researchers have assumed that 40% of participants will be lost during the course of the intervention, leaving a final sample size of 180 (90 in CFs and 90 on other programmes).

The pilot study will combine qualitative and quantitative methods to build knowledge of the mechanisms by which CFs may improve the health and well-being of offenders. The main outcome measure will be quality of life, assessed using the CORE-OM tool[40]. Other outcomes will include: mental health, measured using the WEMBWS positive mental health measure[41]; lifestyle behaviours; use of health and social care services; and reoffending.

The researchers will follow participants beyond their Community Orders and conduct qualitative interviews to understand the support they received and whether or how this has

[38] For a full write-up of the findings go to http://www.theppt.org.uk/documents/Bilderbeck_Farias_2013_J_Psych_Res.pdf

[39] i.e., people dropping out of the process.

[40] http://www.coreims.co.uk/About_Measurement_CORE_Tools.html

[41] http://www.healthscotland.com/documents/1467.aspx

had an impact on their lives. Statistical analysis and triangulation of quantitative and qualitative evidence will be used to develop a model to show how the work of the CF may change the quality of life, health and well-being of offenders. As this is a pilot, it will not be able to draw firm conclusions on differences in effectiveness; however the results will inform the design of a larger study.

**Case Study D: The Brathay Trust: Measuring impact on young people**

The Brathay Trust works with some of the most vulnerable and hard-to-reach young people, and aims to help them develop the confidence, motivation and skills needed to make positive changes in their lives and realise their potential. It delivers community and residential programmes to help young people re-engage with education and avoid or move away from criminal activity.

Brathay is committed to being a 'learning organisation' with an interest in what works and why, which shapes its internal ethos and behaviours such as practice development, strategy, marketing and fundraising.

For all its projects it establishes baselines with young people and measures distance travelled using a range of qualitative, quantitative and creative data collection methods. Brathay regularly records improvement over time amongst the young people it works with. This is encouraging, but doesn't necessarily show that it is Brathay that has made the difference; hence the need to consider the counterfactual, or what would have happened without the service.

As referrals for some projects come from Youth Offending Teams (YOTs), Brathay saw a further opportunity to improve its understanding of impact by developing comparison groups from amongst the YOT's wider client group. Each young person referred to Brathay was matched to another young person by the YOT who was not. Variables such as age, gender and offending history were used to ensure that the match was as close as possible. Analysis was then conducted to compare the two groups by looking at outcomes such as criminal conviction and school exclusions, to see whether referral to Brathay was associated with any differences.

The results are only regarded as indicative because the referral process is not random but based on YOT's judgements (e.g., they may choose young people more suited to Brathay) and because the matching process does not take into account variables such as propensity

to change. However, the analysis is still providing Brathay with a more concrete evidence base from which to think about what it achieves.

**Case Study E: Evaluation of HM Prison Service Enhanced Thinking Skills (ETS) programme**

The main aim of this project was to examine the impact of ETS courses on 'impulsivity' in adult male offenders over the age of 18 in prisons in England and Wales, and to investigate whether changes in levels of impulsivity were reflected in changes in prison behaviour. Impulsivity was chosen as the main outcome measure because of research showing evidence of links between impulsivity and offending (for example Mak (1991), Eysenck and McGurk (1980)).

A Randomised Control Trial (RCT) was used to minimise bias in allocation of participants to groups. Offenders with a priority need to attend a course were assigned to a parallel cohort group prior to the random allocation, and their data were analysed separately. However, it was not possible to measure the association between participation in the ETS course and reoffending as all participants eventually received the intervention—hence there was no comparison group for reoffending analysis.

Analysis of individual cases was also undertaken to investigate evidence of reliable clinical change. A secondary aim was to explore a range of other psychometric measures to evaluate the wider effectiveness of ETS courses, and to examine background factors of offenders and institutional factors to determine which offenders benefit most from ETS programmes, under which conditions.

The study demonstrated positive results with regard to the (short-term) effectiveness of the ETS programme in reducing both self-reported impulsivity and the incidence of prison security reports in adult male offenders. Additionally, the analysis of background factors raised a number of issues relating to which offenders benefited most and how others may be assisted to benefit more.

## Acknowledgements

We are grateful to Margaret Wilson and Alex Sutherland at the Institute of Criminology, University of Cambridge for their input into this guidance. Any errors or admissions are the responsibility of the authors.