Test, Learn, Adapt:

Developing Public Policy with Randomised Controlled Trials

Laura Haynes

Owain Service

Ben Goldacre

David Torgerson

Cabinet Office
Behavioural Insights Team

Dr Laura Haynes is Head of Policy Research at the Behavioural Insights Team. Laura leads the development of randomised controlled trials in a range of policy areas to identify cost-effective policy solutions grounded in behavioural science. Laura has a PhD in Experimental Psychology from the University of Cambridge and is a Visiting Researcher at King's College London.

Owain Service is the Deputy Director of the Behavioural Insights Team. He studied social and political sciences at Cambridge, and has spent most of his career working on public policy in the Prime Minister's Strategy Unit. He has also worked for the Foreign Office in Brussels and for the National Security Secretariat in the UK.

Dr Ben Goldacre trained in medicine and epidemiology and is now a Research Fellow at London School of Hygiene and Tropical Medicine, working on problems in clinical trials. He is the author of Bad Science (4th Estate), a freelance journalist, and has made various documentaries on science, medicine, and policy for BBC Radio 4**.**

Professor David Torgerson is Director of the York Trials Unit. He has a wide interest in randomised trials including those in policy. He has undertaken substantial numbers of clinical trials but also non-medical trials including trials in education; criminal justice and general policy. He has published widely on the methods and methodology of randomised controlled trials.

Acknowledgements

We would like to thank government departments for sharing their recent research involving field trials with us. We'd also like to thank Professor Peter John, Professor Rachel Glennester, Professor Don Green, Dr David Halpern and other members of the Behavioural Insights Team for their feedback on this paper, and also Michael Sanders for editing the document.

Contents

Test

Learn

Adapt

Executive Summary

Randomised controlled trials (RCTs) are the best way of determining whether a policy is working. They are now used extensively in international development, medicine, and business to identify which policy, drug or sales method is most effective. They are also at the heart of the Behavioural Insights Team's methodology.

However, RCTs are not routinely used to test the effectiveness of public policy interventions in the UK. We think that they should be.

What makes RCTs different from other types of evaluation is the introduction of a randomly assigned control group, which enables you to compare the effectiveness of a new intervention against what would have happened if you had changed nothing.

The introduction of a control group eliminates a whole host of biases that normally complicate the evaluation process – for example, if you introduce a new "back to work" scheme, how will you know whether those receiving the extra support might not have found a job anyway?

In the fictitious example below in Figure 1, we can see that those who received the back to work intervention were much more likely to find a job than those who

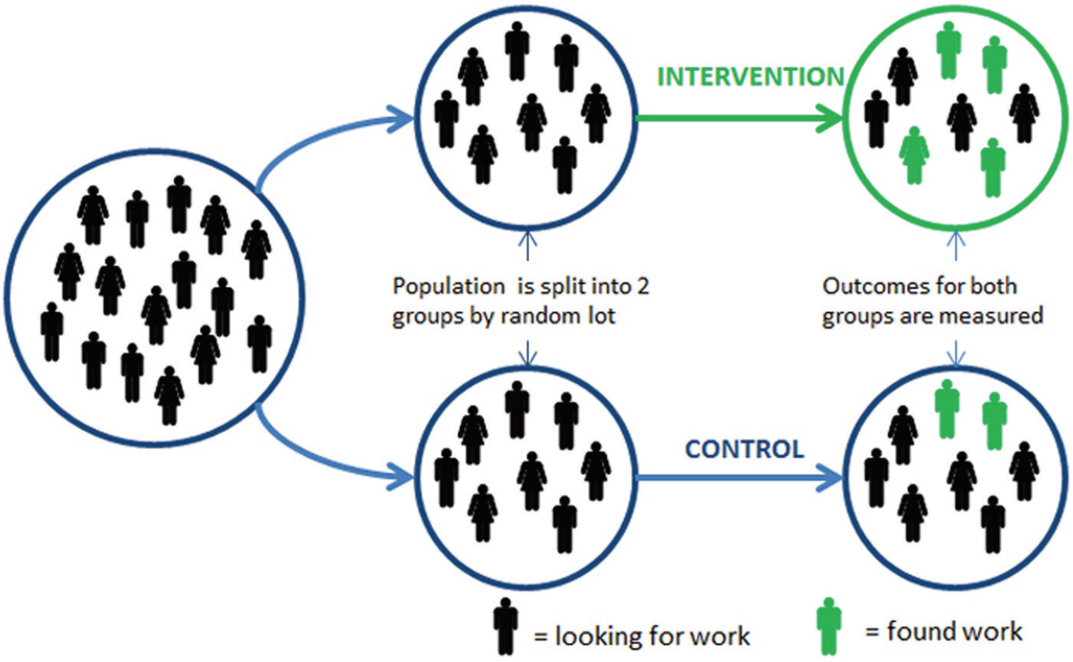

INTERVENTION
Population is split into 2 groups by random lot
Outcomes for both groups are measured
CONTROL
= looking for work
= found work

Figure 1. The basic design of a randomised controlled trial (RCT), illustrated with a test of a new 'back to work' programme.

did not. Because we have a control group, we know that it is the intervention that achieves the effect and not some other factor (such as generally improving economic conditions).

With the right academic and policy support, RCTs can be much cheaper and simpler to put in place than is often supposed. By enabling us to demonstrate just how well a policy is working, RCTs can save money in the long term - they are a powerful tool to help policymakers and practitioners decide which of several policies is the most cost effective, and also which interventions are not as effective as might have been supposed. It is especially important in times of shrinking public sector budgets to be confident that public money is spent on policies shown to deliver value for money.

We have identified nine separate steps that are required to set up any RCT. Many of these steps will be familiar to anyone putting in place a well-designed policy evaluation – for example, the need to be clear, from the outset, about what the policy is seeking to achieve. Some – in particular the need to randomly allocate individuals or institutions to different groups which receive different treatment – are what lend RCTs their power. The nine steps are at the heart of the Behavioural Insights Team's 'test, learn, adapt' methodology, which focuses on understanding better what works and continually improving policy interventions to reflect what we have learnt. They are described in the box adjacent.

Test

1. Identify two or more policy interventions to compare (e.g. old vs new policy; different variations of a policy).

2. Determine the outcome that the policy is intended to influence and how it will be measured in the trial.

3. Decide on the randomisation unit: whether to randomise to intervention and control groups at the level of individuals, institutions (e.g. schools), or geographical areas (e.g. local authorities).

4. Determine how many units (people, institutions, or areas) are required for robust results.

5. Assign each unit to one of the policy interventions, using a robust randomisation method.

6. Introduce the policy interventions to the assigned groups.

Learn

7. Measure the results and determine the impact of the policy interventions.

Adapt

8. Adapt your policy intervention to reflect your findings.

9. Return to Step 1 to continually improve your understanding of what works.

Introduction

Randomised controlled trials (RCTs) are the best way of determining whether a policy is working. They have been used for over 60 years to compare the effectiveness of new medicines.[1] RCTs are increasingly used in international development to compare the cost effectiveness of different interventions for tackling poverty.[2,3] And they are also employed extensively by companies, who want to know which website layout generates more sales. However, they are not yet common practice in most areas of public policy (See Figure 2).

This paper argues that we should and could use RCTs much more extensively in domestic public policy to test the effectiveness of new and existing interventions and variations thereof; to learn what is working and what is not; and to adapt our policies so that they steadily improve and evolve both in terms of quality and effectiveness.

Part I of this paper sets out what an RCT is and why they are important. It addresses many of the common arguments against using RCTs in public policy and argues that trials are

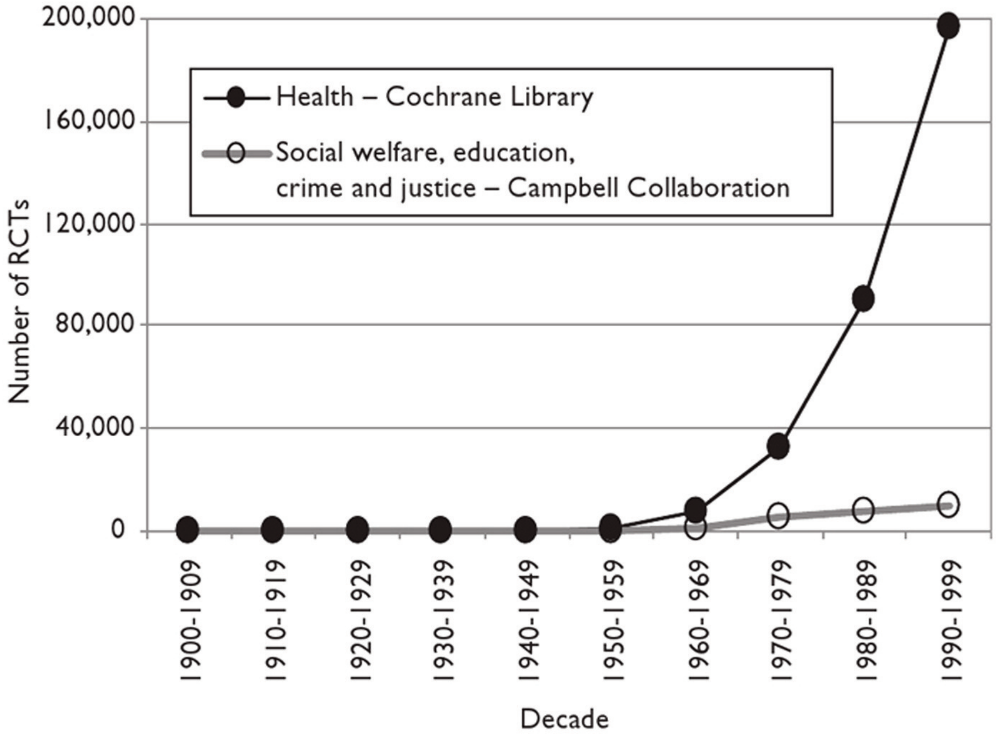


Figure 2. 20th century RCTs in health and in social welfare, education, crime and justice[4]

not as challenging to put in place as is often assumed, and that they can be highly cost-effective ways of evaluating policy outcomes and assessing value for money.

Part II of the paper outlines 9 key steps that any RCT needs to have in place. Many of these steps should be fundamental to any policy initiative, others will require support from academics or centres of expertise within government.

The 'test, learn, adapt' philosophy set out in this paper is at the heart of the way that the Behavioural Insights Team works. We believe that a 'test, learn, adapt' approach has the potential to be used in almost all aspects of public policy:

- **Testing** an intervention means ensuring that you have put in place robust measures that enable you to evaluate the effectiveness or otherwise of the intervention.

- **Learning** is about analysing the outcome of the intervention, so that you can identify 'what works' and whether or not the effect size is great enough to offer good value for money.

- **Adapting** means using this learning to modify the intervention (if necessary), so that we are continually refining the way in which the policy is designed and implemented.

Part I

What is an RCT and why are they important?

What is a Randomised Controlled Trial?

Often we want to know which of two or more interventions is the most effective at attaining a specific, measurable outcome. For example, we might want to compare a new intervention against normal current practice, or compare different levels of "dosage" (e.g. home visits to a teenage expectant mother once a week, or twice a week) against each other.

Conventionally, if you want to evaluate whether an intervention has a benefit, you simply implement it, and then try to observe the outcomes. For example, you might establish a high intensity "back to work" assistance programme, and monitor whether participants come off benefits faster than before the programme was introduced.

However, this approach suffers from a range of drawbacks which make it difficult to be able to identify if it was the intervention that had the effect or some other factor. Principal amongst these are uncontrolled, external factors. If there is strong economic growth, for example, we might expect more people to find employment regardless of our new intervention.

Another, trickier analytical challenge is dealing with so-called "selection bias"; the very people who want to participate in a back to work programme are systematically different to those who do not. They may be more motivated to find work, meaning that any benefits of the new intervention will be exaggerated. There are statistical techniques which people use to try and account for any pre-existing differences between the groups who receive different interventions, but these are always imperfect and can introduce more bias.

Randomised controlled trials get around this problem by ensuring that the individuals or groups of people receiving both interventions are as closely matched as possible. In our "back to work programme" example, this might involve identifying 2000 people who would all be eligible for the new programme and randomly dividing them into two groups of 1000, of which one would get the normal,

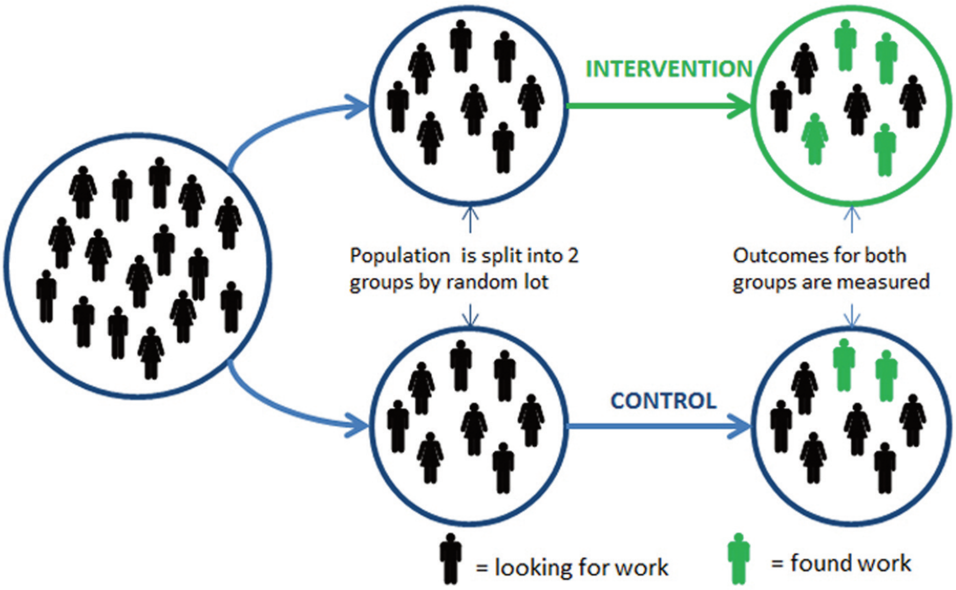

INTERVENTION
Population is split into 2 groups by random lot
Outcomes for both groups are measured
CONTROL
= looking for work
= found work

Figure 3. Illustration of a randomised controlled trial (RCT) to test a new 'back to work' programme (positive outcome).

current intervention and the other would get the new intervention. By randomly assigning people to groups we can eliminate the possibility of external factors affecting the results and demonstrate that any differences between the two groups are solely a result of differences in the interventions they receive.

Part II of this paper describes in more detail how to run a randomised controlled trial, but at the heart of any RCT are a number of key elements. RCTs work by dividing a population into two or more groups by random lot, giving one intervention to one group, the other to another, and measuring the pre-specified outcome for each group. This process is summarised in Figure 3 above.

Let us imagine that we are testing a new "back to work" programme which aims to help job seekers find work. The population being evaluated is divided into two groups by random lot. But only one of these groups is given the new intervention ('the intervention group'), in this case the "back to work" programme. The other group (the 'control group') is given the usual support that a jobseeker would currently be eligible for. In this case, the control group is akin to a

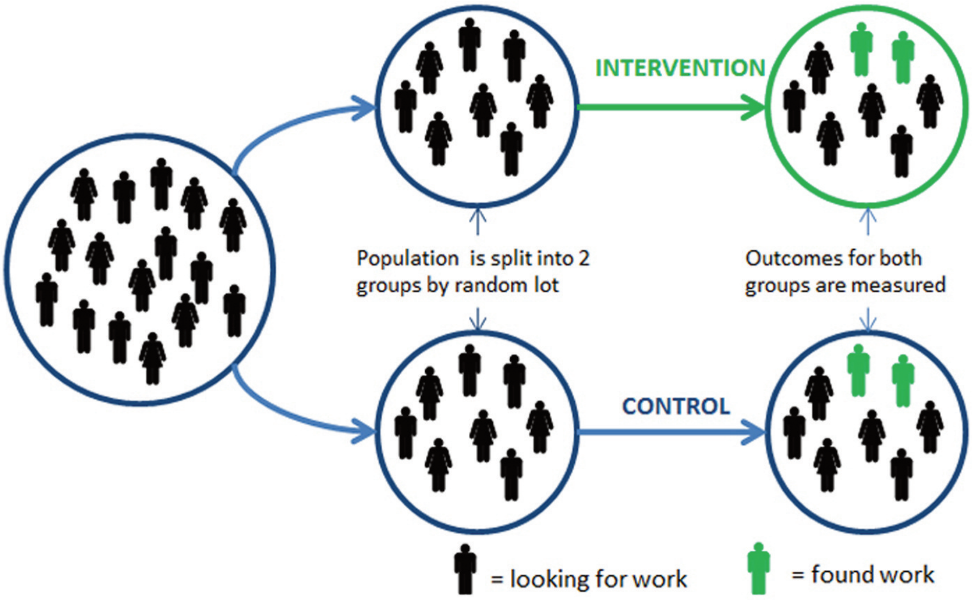

INTERVENTION
Population is split into 2 groups by random lot
Outcomes for both groups are measured
CONTROL
= looking for work
= found work

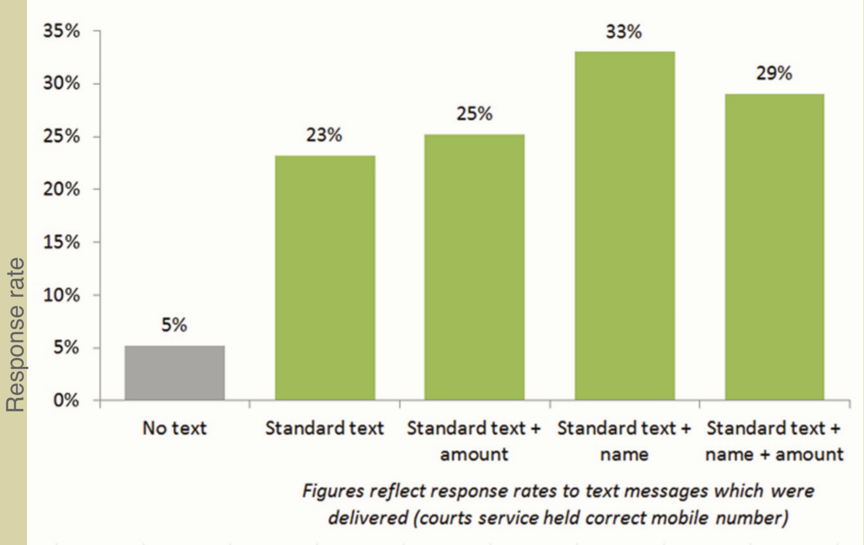
Figure 4. Illustration of a randomised controlled trial (RCT) to test a new 'back to work' programme (neutral outcome).

Box 1: Demonstrating the impact of text messaging on fine repayments

The Courts Service and the Behavioural Insights Team wanted to test whether or not sending text messages to people who had failed to pay their court fines would encourage them to pay prior to a bailiff being sent to their homes. The way this question was answered is a clear example of the “test, learn, adapt” approach, and the concurrent testing of multiple variations to find out what works best.

In the initial trial, individuals were randomly allocated to five different groups. Some were sent no text message (control group), while others (intervention groups) were sent either a standard reminder text or a more personalised message (including the name of the recipient, the amount owed, or both).

The trial showed that text message prompts can be highly effective (Figure 5).



Figures reflect response rates to text messages which were delivered (courts service held correct mobile number)

Figure 5. Initial trial: repayment rates by individuals (N=1,054)

A second trial was conducted using a larger sample (N=3,633) to determine which aspects of personalised messages were instrumental to increasing payment rates. The pattern of results was very similar to the first trial. However, the second trial enabled us to be confident not only that people were more likely to make a payment on their overdue fine if they received a text message containing their name, but that the average value of fine repayments went up by over 30%.

The two trials were conducted at very low cost: as the outcome data was already being collected by the Courts Service, the only cost was the time for team members to set up the trial. If rolled out nationally, personalised text message reminders would improve collection of unpaid fines; simply sending a personalised rather than a standard text is estimated to bring in over £3 million annually. The savings from personalised texts are many times higher than not sending any text reminder at all. In addition to these financial savings, the Courts Service estimates that sending personalised text reminders could reduce the need for up to 150,000 bailiff interventions annually.

placebo condition in a clinical drug trial.

In the example in Figure 3, jobseekers who have found full time work 6 months into the trial are coloured green. The trial shows that many more of the individuals in the new "back to work" programme are now in work compared to those in the control group.

It is important to note that two stick figures in the control group have also found work, perhaps having benefited from the normal jobseeker support provided to all those on benefits.

If the new "back to work" programme was no better than the current service provided to jobseekers, we would have seen a similar pattern in both the intervention group and the control group receiving the current service. This is illustrated best in Figure 4, which shows a different set of results for our programme.

Here, the results of the trial demonstrate that the new, expensive "back to work" programme is no better than current practice. If there had been no control group, we might have seen people getting jobs after taking part in the new "back to work" programme, and wrongly concluded that they had done so because of the programme itself. This might have led us to roll out the new, expensive (and ineffective) intervention. A mistake like this was avoided by the DWP in a real life RCT looking at the cost-effectiveness of different types of interventions (see Box 2).

Wherever there is the potential for external factors to affect the outcomes of a policy, it is always worth considering using an RCT to test the effectiveness of the intervention before implementing it in the whole population. When we do not, it is easy to confuse changes that might have occurred anyway with the impact of a particular intervention.

Our fictitious "back to work" example assumes that we are interested principally

Box 2. Using RCTs to know what really works to people get back into employment.

In 2003, the Department for Work and Pensions (DWP) conducted an RCT to examine the impact of three new programmes on Incapacity Benefit claimants: support at work, support focused on their individual health needs, or both.[5,6] The extra support cost £1400 on average, but the trial found no benefit over the standard support that was already available. The RCT ultimately saved the taxpayer many millions of pounds as it provided unambiguous evidence that the costly additional support was not having the intended effect.

More recently the DWP was keen to explore whether the intensity of the signing-on process required of jobseekers on benefits could be reduced without worsening outcomes.

In a trial involving over 60,000 people, the usual fortnightly signing-on process was compared against several others which were less resource intensive (e.g. signing-on by telephone, less frequently). All of the alternatives to the status quo tested in trials large enough to show reliable effects were found to increase the time people took to find work.[7] As a result, despite other changes to the benefits system, DWP policy continues to require people to sign on a fortnightly basis.

in understanding which of two large-scale interventions is working most effectively. In many cases, an RCT is not just interested in the headline policy issue. Instead, it may be used to compare several different ways of implementing smaller aspects of the policy.

As many of the other examples set out in this paper show, one of the great things about randomised controlled trials is that they also allow you to test the effectiveness of particular aspects of a wider programme. Testing small parts of a programme enables policy makers to continually refine policy, honing in on the particular aspect of the intervention which is having the greatest impact.

Regardless of whether we are comparing two large scale interventions or smaller aspects of a single policy, the same basic principles of an RCT hold true: by comparing two identical groups, chosen at random, we can control for a whole range of factors that enable us to understand what is working and what is not.

Box 3: The link between theories of growth, innovation, and RCTs

The growing interest in the use of RCTs as an important tool of policymaking and practice resonate with broader currents of thinking. When money is short it is essential to make sure that it is being spent on approaches that work, and even small marginal improvements in cost effectiveness are precious. RCT's are an extremely powerful tool to pinpoint cost-effectiveness – and flush out low-value spend.

These methods also resonate strongly with emerging views on social and economic progress. Many leading thinkers have concluded that in complex systems, from biological ecosystems to modern economies, much progress – if not most – occurs through a process of trial and error.[8] Economies and ecosystems that become too dominated by a narrow a range of practices, species or companies are more vulnerable to failure than more diverse systems.[9,10] Similarly, such thinkers tend to be sceptical about the ability of even the wisest experts and leaders to offer a comprehensive strategy or masterplan detailing 'the' best practice or answer on the ground (certainly on a universal basis). Instead they urge the deliberate nurturing of variation coupled with systems, or dynamics, that squeeze out less effective variations and reward and expand those variations that seem to work better.

The practical expression of this thinking includes the drive for greater devolution of policy-making, and the harnessing of markets to deliver goods and services. Encouraging variation needs to be matched by mechanisms that identify and nurture successful innovations. This includes sharpening transparency and feedback loops in consumer markets and public services, noting that these lead to the selective expansion of better provision and often the growth of smaller, independent provision.[11] In public services, and where markets and payment by results may be inappropriate, RCTs and multi-arm trials may play a powerful role here, especially where these results are widely reported and applied.

The case for RCTs: debunking some myths

There are many fields in which randomised trials are now common practice, and where failing to do them would be regarded as bizarre, or even reckless. RCTs are the universal means of assessing which of two medical treatments works best, whether it is a new drug compared with the current best treatment, two different forms of cancer surgery, or even two different compression stockings. This was not always the case: when trials were first introduced in medicine, they were strongly resisted by some clinicians, many of whom believed that their personal expert judgement was sufficient to decide whether a particular treatment was effective.

RCTs are also increasingly being used to investigate the effectiveness and value for money of various different international development programmes (see Box 4). In business, when companies want to find out which of two webpage designs will encourage the most "click-throughs" and sales, it is common to randomly assign website visitors to one of several website designs, and then track their clicks and purchasing behaviour (see Box 5).

But while there are some good examples of policymakers using RCTs in the UK, they are still not in widespread use. This may partly be due to a lack of awareness, but there are also many misunderstandings about RCTs, which lead to them being inappropriately rejected.

Here we go through each of these myths in turn, addressing the incorrect assumption that RCTs are always

Box 4. Using RCTs to improve educational outcomes in India

One of the areas of rapid growth in the use of RCTs in recent years has been in international development. Numerous trials have been conducted to determine how best to tackle poverty in the developing world, from how to tackle low crop yields, to how encourage use of mosquito nets, ensure teachers turn up at class, foster entrepreneurship, and increase vaccination rates.

For example, effort in recent decades to make education universally available in developing countries led to improved school enrolment and attendance. However, the quality of education available to children from poor background remains an issue: one 2005 survey across India indicated that over 40% of children under 12 could not read a simple paragraph, and 50% couldn't perform a simple subtraction.

In partnership with an education NGO, US researchers conducted an RCT to determine whether a low cost, in school remedial education programme could improve school outcomes in India. Almost 200 schools were randomly allocated to receive a tutor for either their third or fourth grade. The impact of the programme was ascertained by comparing grade 3 outcomes for those schools with and without grade 3 tutors.

The tutors were women from the local community who were paid a fraction of a teacher's salary, and they worked separately with groups of children who were falling behind their peers for half of the school day. Results indicated that the remedial programme significantly improved test scores, particularly in maths.[12] The programme was judged so successful (and cost effective relative to other programmes to improve school performance) that it has been scaled up across India.

difficult, costly, unethical, or unnecessary. We argue that it is dangerous to be overconfident in assuming that interventions are effective, and that RCTs play a vital role in demonstrating not only the effectiveness of an intervention, but also value for money.

Box 5. Using RCTs to improve business performance

Many companies are increasingly using RCT designs to test consumer responses to different presentations of their products online. Little of this information is publicly available, but it is well known that companies such as Amazon and eBay use routine web traffic on their sites to test out what works best to drive purchases. For example, some customers might view a particular configuration of a webpage, while others will view a different one. By tracking the "click-throughs" and purchasing behaviour of customers who view the different versions of the website, companies can tweak web page designs to maximise profits. A few examples are provided below.

During the recent Wikipedia fund-raising drive, a picture of the founder, Jimmy Wales, appeared in the donations advert at the top of the page: this was the result of a series of trials comparing different designs of advert, delivering them randomly to website visitors, and monitoring whether or not they donated.

Netflix is a company that offers online movie streaming, and typically runs several user experience experiments simultaneously. When they were trialling the "Netflix Screening Room", a new way to preview movies, they produced four different versions of the service. These were each rolled out to four groups of 20,000 subscribers, and a control group received the normal Netflix service. Users were then monitored to see if they watched more films as a result.[13]

Delta Airlines have also used experimentation to improve their website design. In 2006, while increasing numbers of people were booking their travel online, web traffic to Delta Airlines' website was failing to generate the anticipated number of bookings. Almost 50% of the visitors to their website were dropping off before completing the booking process: after selecting their flight, potential customers often abandoned the booking when they reached the web page requiring input of their personal information (name, address, card details),

Rather than changing the entire website, Delta focused on making changes to the specific pages which failed to convert potential customers into sales. Numerous variations were tested online, randomly assigning customers to different versions of the webpages. Delta discovered that by removing detailed instructions at the top of the page requesting their personal information, customers were much more likely to complete the booking process. As a result of implementing this and other subtle design changes identified during the testing process, conversion rates to ticket sales have improved by 5%[14], a small but highly valuable change.

1. We don't necessarily know 'what works'

Policymakers and practitioners often feel they have a good understanding of what interventions are likely to work, and use these beliefs to devise policy. Even if there are good grounds for believing a policy will be effective, an RCT is still worthwhile to quantify the benefit as accurately as possible. A trial can also help to demonstrate which aspects of a programme are having the greatest effect, and how it could be further improved. For example, if we were implementing a new programme for entrepreneurs based on start-up funding, it would be useful to know whether doubling the amount of available funding has a significant effect on success or makes no difference.

We should also recognise that confident predictions about policy made by experts often turn out to be incorrect. RCTs have demonstrated that interventions which were designed to be effective were in fact not (see Box 2). They have also shown that interventions about which there was initial scepticism were ultimately worthwhile. For example, when the Behavioural Insights Team and the Courts Service looked at whether text messaging might encourage people to pay their court fines, few predicted at the outset that a personalised text would increase repayment rates and amounts so significantly (see Box 1).

But there are also countless examples of RCTs that have overturned traditional assumptions about what works, and showed us that interventions believed to be effective were, in reality, harmful. The steroid injection (see Box 6) case is a powerful example of how apparently sound assumptions do not necessarily hold true when finally tested. Similarly, the Scared Straight programme, which exposes young people to the realities of a life of crime, is a good example of a well-intentioned policy intervention with an apparently sound evidence base, but which RCTs have shown adverse effects (see Box 7). RCTs are the best method we have for avoiding these mistakes, by giving policymakers and practitioners robust evidence of the effectiveness of a policy intervention, and ensuring that we know what would have happened in the absence of the intervention.

2. RCTs don't have to cost a lot of money

The costs of an RCT depend on how it is designed: with planning, they can be cheaper than other forms of evaluation. This is especially true when a service is already being delivered, and when outcome data is already being collected from routine monitoring systems, as in many parts of the public sector. In contrast to trials in medicine, a public policy trial will not necessarily require us to recruit participants outside of normal practice or to put new systems in place to deliver interventions or monitor outcomes.

The Behavioural Insights Team has worked with a range of different government departments to run trials at little additional cost to the time of team members. For example, in trials the team has run with local authorities, HMRC, DVLA, and the Courts Service (and is about to run with Job Centre Plus), the data is already being

Box 6. Steroids for head injury: saving lives, or killing people?

For several decades, adults with severe head injury were treated using steroid injections. This made perfect sense in principle: steroids reduce swelling, and it was believed that swelling inside the skull killed people with head injuries, by crushing their brain. However, these assumptions were not subject to proper tests for some time.

Then, a decade ago, this assumption was tested in a randomised trial. The study was controversial, and many opposed it, because they thought they already knew that steroids were effective. In fact, when the results were published in 2005[15], they showed that people receiving steroid injections were more likely to die: this routine treatment had been killing people, and in large numbers, because head injuries are so common. These results were so extreme that the trial had to be stopped early, to avoid any additional harm being caused.

This is a particularly dramatic example of why fair tests of new and existing interventions are important: without them, we can inflict harm unintentionally, without ever knowing it; and when new interventions become common practice without good evidence, then there can be resistance to testing them in the future.

routinely collected and processes are already in place to deliver interventions, whether it is a letter, or a fine, or an advisory service for unemployed people.

When considering the additional resources that might be required to run an RCT, we should remember that they are often the best way to establish if a programme offers good value for money. In some cases, a trial may lead to us to conclude that a programme is too expensive to roll out, if the extra benefits of the intervention are negligible. In others, a trial may demonstrate that a programme delivers excellent value for money, and so should be rolled out more widely.

By demonstrating how much more or less effective the intervention was than the status quo, policymakers can determine whether the cost of the intervention justifies the benefits. Rather than considering how much an RCT costs to run, then, it might be more appropriate to ask: *what are the costs of not doing an RCT?*[16]

3. There are ethical advantages to using RCTs

Sometimes people object to RCTs in public policy on the grounds that it is unethical to withhold a new intervention from people who could benefit from it. This is particularly the case where additional money is being spent on programmes which might improve the health, wealth, or educational attainment of one group.

It is true to say that it can be challenging to withhold a treatment or intervention from someone that we believe might benefit from it. This paper does not argue that we should do so when we know that an intervention is already proven to be beneficial.

However, we do argue that we need to be clear about the limits of our knowledge and that we will not be certain of the effectiveness of an intervention until it is tested robustly.

Sometimes interventions which were believed to be effective turned out to be ineffective or even actively harmful (see Boxes 6 and 7). This can even be the case with policies that we might intuitively think will be guaranteed to work. For example, incentives have been used to encourage adult learners to attend literacy classes, but when an RCT of this policy was conducted, it was found that participants receiving incentives attended approximately 2 fewer classes per term than the non-incentive group.[20]

In this trial, using small incentives not only wasted resources, it actively reduced class attendance. Withholding the intervention was better than giving it out, and if a trial had never been conducted, we could have done harm to adult learners, with the best intentions, and without ever knowing that we were doing so.

It is also worth noting that policies are often rolled out slowly, on a staggered basis, with some regions "going early", and these phased introductions are not generally regarded as unethical. The delivery of the Sure Start programme is an example of this.

If anything, a phased introduction in the context of an RCT is more ethical, because it generates new high quality information that may help to demonstrate that an intervention is cost effective.

Box 7: The Scared Straight Programme: Deterring juvenile offenders, or encouraging them?

"Scared Straight" is a programme developed in the US to deter juvenile delinquents and at-risk children from a criminal behaviour. The programme exposed children to the frightening realities of leading a life of crime, through interactions with serious criminals in custody.

The theory was that these children would be less likely to engage in criminal behaviour if they were made aware of the serious consequences. Several early studies, which looked at the criminal behaviours of participants before and after the programme, seemed to support these assumptions.[17] Success rates were reported as being as high as 94%, and the programme was adopted in several countries, including the UK.

None of these evaluations had a control group showing what would have happened to these participants if they had not participated in the programme. Several RCTs set out to rectify this problem. A meta-analysis of 7 US trials which randomly assigned half of the sample of at-risk children to the programme and found that "Scared Straight" in fact led to higher rates of offending behaviour: "doing nothing would have been better than exposing juveniles to the program".[18] Recent analyses suggest that the costs associated with the programme (largely related to the increase in reoffending rates) were over 30 times higher than the benefits, meaning that "Scared Straight" programmes cost the taxpayer a significant amount of money and actively increased crime.[19]

4. RCTs do not have to be complicated or difficult to run

RCTs in their simplest form are very straightforward to run. However there are some hidden pitfalls which mean that some expert support is advisable at the outset.

Some of these pitfalls are set out in the next chapter, but they are no greater than those faced in any other form of outcome evaluation, and can be overcome with the right support. This might involve, for example, making contact with the Behavioural Insights Team. We can advise on trial design, put policy makers in touch with academics who have experience of running RCTs, and can help to guide the design of a trial. Very often, academics will be happy to assist in a project which will provide them with new evidence in an area of interest to their research, or the prospect of a published academic paper.

The initial effort to build in randomisation, and clearly define outcomes before a pilot is initiated, is often time well spent. If an RCT is not run, then any attempt to try and evaluate the impact of an intervention will be difficult, expensive, and biased - using complex models will be required to try and disentangle observed effects which could have multiple external causes. It is much more efficient to put a smaller amount of effort into the design of an RCT before a policy is implemented.

Box 8: Family Nurse Partnership: building in rigorous evaluation to a wider roll out.

The Family Nurse Partnership (FNP) is a preventative programme for vulnerable first time mothers. Developed in the US, it involves structured, intensive home visits by specially trained nurses from early pregnancy until the child is two years of age. Several US RCTs[21] have shown significant benefits for disadvantaged young families and substantial cost savings. For example, FNP children have better socio-emotional development and educational achievement and are less likely to be involved in crime. Mothers have fewer subsequent pregnancies and greater intervals between births, are more likely to be employed and less likely to be involved in crime.

FNP has been offered in the UK since 2007, often through Sure Start Children's Centres, and the Department of Health has committed to doubling the number of young mothers receiving support through this programme to 13,000 (at any one time) in 2015. Meanwhile, the Department is funding an RCT evaluation of the programme, to assess whether FNP benefits families over and above universal services and offers value for money. It involves 18 sites across the UK, and approximately 1650 women, the largest trial to date of FNP. Reporting in 2013, outcomes measures include smoking during pregnancy, breastfeeding, admissions to hospital for injuries and ingestions, further pregnancies and child development at age 2.

PART II

Conducting an RCT: 9 key steps

How do you conduct a Randomised Controlled Trial?

Part I of this paper makes the case for using RCTs in public policy. Part II of this paper is about how conduct an RCT. It does not attempt to be comprehensive. Rather, it outlines the necessary steps that any RCT should go through and points to those areas in which a policy maker may wish to seek out more expert advice.

We have identified nine separate steps that any RCT will need to put in place. Many of these nine steps will be familiar to anyone putting in place a well-designed policy evaluation – for example, the need to be clear, from the outset, what the policy is seeking to achieve.

Several, however, will be less familiar, in particular the need to randomly allocate the intervention being tested to different intervention groups. These are summarised below and set out in more detail in the sections that follow.

Test

1. Identify two or more policy interventions to compare (e.g. old vs new policy; different variations of a policy).
2. Determine the outcome that the policy is intended to influence and how it will be measured in the trial.
3. Decide on the randomisation unit: whether to randomise to intervention and control groups at the level of individuals, institutions (e.g. schools), or geographical areas (e.g. local authorities).
4. Determine how many units (people, institutions, or areas) are required for robust results.
5. Assign each unit to one of the policy interventions, using a robust randomisation method.
6. Introduce the policy interventions to the assigned groups.

Learn

7. Measure the results and determine the impact of the policy interventions.

Adapt

8. Adapt your policy intervention to reflect your findings.
9. Return to Step 1 to continually improve your understanding of what works

Test

Step 1: Identify two or more policy interventions to compare

RCTs are conducted when there is uncertainty about which is the best of two or more interventions, and they work by comparing these interventions against each other. Often, trials are conducted to compare a new intervention against current practice. The new intervention might be a small change, or a set of small changes to current practice; or it could be a whole new approach which is proving to be successful in a different country or context, or that has sound theoretical backing.

Before designing an RCT, it is important to consider what is currently known about the effectiveness of the intervention you are proposing to test. It may be, for example, that RCTs have already been conducted in similar contexts showing the measure to be effective, or ineffective. Existing research may also help to develop the policy intervention itself. A good starting point are the Campbell Collaboration archives[22], which support policymakers and practitioners by summarising existing evidence on social policy interventions.

It is also important that trials are conducted on the very same intervention that would be rolled out if the trial was successful. Often there is a temptation to run an RCT using an ideal, perfect policy intervention, which is so expensive that it could never be rolled out nationwide. Even if we did have the money, such an RCT would be uninformative, because the results would not generalise to the real-world policy implementation.

We need to be sure that the results of our trial will reflect what can be achieved should the policy be found to be effective and then rolled out more widely. In order for findings to be generalisable, and relevant to the whole country, the intervention must be representative, as should the eagerness with which practitioners deliver it, and the way data is collected.

The Behavioural Insights Team, in conducting public policy RCTs, will usually spend a period of time working with front-line organisations to both understand what is likely to be feasible, and to learn from staff who themselves might have developed potentially effective but untested new methods for achieving public policy outcomes.

Box 9. Comparing different policy options & testing small variations on a policy

An RCT is not necessarily a test between doing something and doing nothing. Many interventions might be expected to do better than nothing at all. Instead, trials can be used to establish which of a number of policy intervention options is best.

In some cases, we might be interested in answering big questions about which policy option is most appropriate. For example, imagine we had the money to upgrade the IT facilities in all secondary schools, or pay for more teachers, but not both. We might run a 3 arm trial (see figure 6), with a control group (a number of schools continuing with current IT and the same number of teachers) and two intervention groups (schools who received an IT upgrade or more teachers). This would enable us to determine whether either new policy option was effective, and which offered the best value for money.

In other cases, we might be interested in answering more subtle questions about a particular policy, such as which minor variation in delivery leads to the best outcomes. For example, imagine that we are planning on making some changes to the food served in school canteens. We might already be bringing in some healthier food options with the intention of improving children's diets. However, we know that presentation matters, and we aren't sure how best to lay out the food options to encourage the healthiest eating. We might run a multi arm trial, varying the way in which the food is laid out (the order of salads and hot foods, the size of the ladles and plates, etc).

Opportunities to fine tune policies often arise when we are about to make changes – it is an ideal time to test out a few minor variations to ensure that the changes we finally institute are made to best effect.

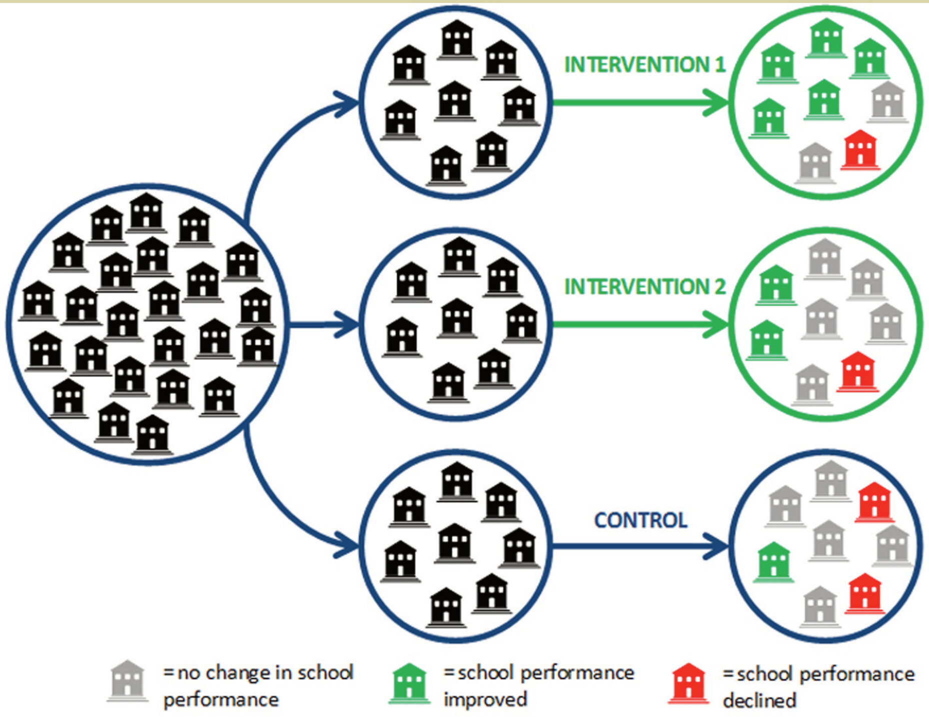

INTERVENTION 1
INTERVENTION 2
CONTROL
= no change in school performance
= school performance improved
= school performance declined

Figure 6. The design of a hypothetical multi-arm RCT testing whether upgrading schools' IT facilities (intervention 1) or employing more teachers (intervention 2) improves the school's academic performance.

Step 2: Define the outcome that the policy is intended to influence and how it will be measured in the trial

It is critical in any trial that we define at the outset exactly what outcome we are trying to achieve and how we will measure it. For example, in the context of educational policy, an outcome measure might be examination results. For policies related to domestic energy efficiency, an outcome measure could be household energy consumption.

It is important to be specific about how and when the outcomes will be measured at the design stage of the trial, and to stick with these pre-specified outcomes at the analysis stage. It is also critical to ensure that the way outcomes are measured for all the groups is exactly the same – both in terms of the process of measurement and the standards applied.

Pre-specifying outcome measures does not just make good practical sense. There are also good scientific reasons why it is crucial to the success of a well-run RCT. This is because, over the course of time, there will always be random fluctuations in routinely collected data. At the end of the trial, there may be a lot of data on a lot of different things, and when there is so much data, it is inevitable that some numbers will improve – or worsen – simply through random variation over time.

Box 10: Taking advantage of natural opportunities for RCTs

Sometimes constraints on policy delivery provide the ideal context for a policy trial. For example, financial constraints and/or practicalities may mean that a staggered roll-out is the preferred option. As long as there is a facility to monitor outcomes in all the areas which will eventually receive the policy intervention, and there is a willingness to randomly decide which area goes first, a staggered policy roll out can be exploited to run a ‘stepped-wedge’ design trial.

For example, the probation service in the Durham area wanted to test out a new approach to delivering the probation service. Resource constraints precluded all 6 probation centres receiving the new guidance and training at the same time. The fairest, and scientifically most robust, approach was to randomly assign the 6 centres to a position in a waiting list. All centres eventually received the training but because random allocation rather than administrative convenience determined when each centre received the training, a robust evaluation of the effects of the new service on reoffending rates could be conducted.[23]

Whenever such random variation occurs, it may be tempting to pick out some numbers that have improved, simply by chance, and view those as evidence of success. However, doing this breaks the assumptions of the statistical tests used to analyse data, because we are giving ourselves too many chances to find a positive result. The temptation to over-interpret data, and ascribe meaning to random variation, is avoided by pre-specifying outcomes. Statistical tests can then be meaningfully used to analyse how much of the variation is simply due to chance.

When deciding on an outcome measure, it is also important to identify an outcome that you really care about, or as close as you can get to it, rather than a procedural measure that is halfway there. For example, in a trial to see whether probation officers referring to alcohol services can reduce re-offending, you might measure: alcohol service referrals, alcohol service attendances, alcohol intake by questionnaire, or re-offending.

In this case re-offending is the outcome we care about the most, but that data might be harder to collect and any benefit on offending might take years to become apparent. Because of this, you could consider measuring alcohol service attendance, as a "surrogate outcome" for the real outcome of offending behaviour. Alternatively, you might measure both: service attendance, to give interim findings; and then long-term follow-up results on offending 24 months later. "Referrals by probation officers" would be the easiest thing to measure, and although immediate, it is not ultimately very informative if re-offending is what we really care about. See Box 11 for an example.

The question of which outcome measure to use often benefits from collaborative discussion between academics (who know what would work best technically in a trial) and policymakers (who know what kind of data is conveniently available, and what it might cost to collect).

Box 11: The case for (and against) using surrogate outcomes.

A surrogate outcome is one which is a proxy for the true outcome of interest; for example, reconviction rates are used as a surrogate for reoffending rates, because they are far easier to measure (as people might be never caught for the crimes they commit). The case for using a surrogate outcome is strongest where there is good evidence that it is a strong predictor of the ultimate outcome of interest. Unfortunately, using self-reported measures of behaviour change, while easy to measure, can be a poor index of actual behavioural change. Due to "social desirability" biases, people may be motivated to over-report, for example, the amount of exercise they do, after they have taken part in a "get fit" programme.

If surrogate outcomes are needed because the final outcomes are very long term, it is always worthwhile following up these long term outcomes to verify the interim results. There are numerous cases in clinical medicine where initial trials using surrogate outcomes were misleading. For example, offering patients with osteoporosis fluoride treatment was thought to be effective as it led to increased bone density. As one of the key clinical indicators of osteoporosis, bone density was judged an appropriate surrogate outcome. However, it has been demonstrated that fluoride treatment in fact leads to an increase in some types of fractures, the ultimate outcome osteoporotic patients are keen to avoid.[.24]

Step 3: Decide on the randomisation unit

After deciding what outcome we are going to measure (Step 2), we need to decide who or what we are going to randomise. This is known as the randomisation unit.

The randomisation unit is most often individual people, for example when individuals are randomly assigned to receive one of two medical treatments, or one of two educational programmes. However, the randomisation unit can also be a group of people centred around an institution, especially if the intervention is something that is best delivered to a group. For example, whole schools might be randomly assigned to deliver a new teaching method, or the current one; whole Job Centres might be randomly assigned to offer a new training programme, or the current one. Lastly, the randomisation unit could be a whole geographical area: for example, local authorities might be randomly assigned to deliver one of two new health prevention programmes or different methods of waste recycling (see Box 12).

At the end of the trial, outcomes can be measured in individuals, or for the whole randomisation unit, depending on what is practical, and most accurate. For example, although whole classes might be randomly assigned to receive different teaching methods, the learning outcomes of individual students can be assessed when calculating the results, for greater accuracy.

Box 12. Capitalising on local variations in policy

Local authorities are well placed to test new policies in the field. By collaborating with other local authorities to trial different policies, or by randomly assigning different streets or regions to different interventions, local authorities can use the RCT methodology to determine what works.

An example of this approach is the trial conducted by the local authority of North Trafford to compare different methods of promoting waste recycling. The randomisation unit in this trial was "whole streets". Half of the streets in one part of the local authority were randomly assigned to be canvassed to encourage them to recycle their waste. Recycling rates were higher in this group, compared with the almost 3000 households who did not receive the canvassing.

The increase over the short term was 5%, and the academic partners judged that the canvassing campaign cost around £24 for every additional household who started recycling.[25] Based on this information, the local authority was then in a position to determine whether the reduced landfill costs associated with the canvassing campaign could justify the costs of offering it more widely.

The question as to whether the randomisation unit should be individuals, institutions or areas will usually depend upon practical considerations. In clinical trials, for example, it will usually be possible to give different individuals either a placebo or the drug which is being tested. But in public policy trials, it may not always be possible to do so. Below we consider two examples of different ways in which the Behavioural Insights Team has decided upon what the randomisation unit should be:

- **Individual**: When considering different messages in tax letters, it is obviously possible to send different letters out to different individuals, so the randomisation unit was individual tax debtors.
- **Institution**: When running a trial on supporting people to get into work in Job Centres, it is not possible to randomly assign different interventions to different job seekers, so the randomisation unit will be Job Centre teams (i.e. teams of the advisors who help the job seekers).

As with other steps, it will be useful to discuss the randomisation unit with an academic advisor. It will also be important to consider how the decision to choose a particular unit interacts with other considerations. Most importantly, it will affect how many people will need to be involved in the trial: having institutions or areas as your unit of study will nearly always mean that a larger sample of individuals is required, and special methods of analysis are also needed.

There can also be other considerations: for example, in an evaluation of attendance incentives for adult education classes, the researchers chose to randomise whole classes, even though it would have been possible to randomise individual attendees. This was to avoid resentful demoralisation from those in the group without incentives, who would see that other learners in their class were receiving an incentive and they were not. This may have negatively affected their attendance rate to the classes, and we might have seen an effect due

Box 13. When the randomisation unit should be groups rather than individuals

Worms like hookworm infect almost one quarter of the world's population, mostly in developing countries. It is a common cause of school absence, and US researchers collaborated with the US Ministry of Health to determine whether offering children deworming treatment would reduce school absenteeism.

An RCT was conducted in which entire schools either received mass deworming treatment or continued as usual without it. In this case, individual randomisation would have been inappropriate – if randomisation had occurred within schools, such that some pupils were dewormed and others were not, the likelihood that the control participants would contract an infection may have been artificially reduced by the fact their peers were worm-free.

Seventy-five primary schools in rural Kenya took part in the study, which demonstrated that the deworming programme reduced absenteeism by one quarter.[26] Increases in school attendance were particularly marked in the youngest children. This study demonstrated that an additional year of school attendance could be achieved by deworming at cost of $3.50 per student, representing a highly cost-effective method to increase school participation (other programmes, such as free school uniforms, cost over $100 per student to deliver similar effects).[3]

to this problem rather than due to the incentive.

In addition, it is crucially important that individuals are recruited to the study before the randomisation is done, otherwise the trial ceases to be robust.

For example, if the people running a trial know which group a potential participant would be allocated to, before that participant is formally recruited into the study, then this may affect the decision to recruit them at all. A researcher or frontline staff member who believes passionately in the new intervention may choose – maybe unconsciously – not to recruit participants who they believe are “no hopers” into the new intervention group. This would mean that the participants in each “random” group were no longer representative. This kind of problem can be avoided by simply ensuring that participants are formally recruited into the trial first, and then randomised afterwards.

Step 4: Determine how many units are required for robust results

To draw policy conclusions from an RCT, the trial must be conducted with a sufficient sample size. If the sample size is large enough, we can be sure that the effect of our intervention is unlikely to be due to chance.

If we have decided that the randomisation unit will be institutions or areas, it is very likely that we will need a larger number of people in the trial than if we had decided to randomise by individual. Simple preliminary “power calculations” will help determine how many units (individuals, institutions etc.) should be included in the policy intervention and control groups. We recommend working with academics who have experience in RCTs to ensure this key technical calculation is done correctly.

If your policy intervention delivers a huge benefit (a large “effect size”), you will be able to detect this using a trial with a relatively small sample size. Detecting more subtle differences (small effect sizes) between interventions will require larger numbers of participants, so it important from the outset not to be overly optimistic about the likely success of an intervention. Many interventions - if not most - have relatively small effects.

As an example of how many participants are needed for a trial: if we randomly allocated 800 people into two groups of 400 each this would give us about an 8 out of 10 chance of seeing a difference of 10%, if such a difference existed.

For example, imagine that the government wants to encourage people to vote, and wants to test the effectiveness of sending text messages to registered voters on the morning of an election to remind them. They choose 800 voters to observe: 400 in the control group who will receive no extra reminder, and 400 in the treatment group, who will receive text messages. If turnout is 50% in the control group, with a sample of this size we would have an 80% chance of seeing a change from 50% to 60% (a 10 percentage point change). If we wanted to detect a smaller difference, we would need larger sample sizes.

Some consideration should be given to how much it costs to recruit each additional person and the impact (effect size and potential cost savings) of the intervention that is being measured. Sometimes detecting even a modest difference is very useful, particularly if the intervention itself costs little or nothing. For example, if we are changing the style or content of a letter to encourage the prompt payment of tax, then the additional cost is very small, as postage costs are incurred anyway and we are already collecting the outcome data (in this case, payment dates). In contrast, if we wanted to increase the proportion of people who are on Job Seekers' Allowance getting a full time job by giving them one to one job related counselling, then this is relatively expensive, and we would hope to see a commensurately larger effect for it to be worthwhile running a trial. However, even for expensive interventions, if hypothesised impacts are small in terms of effect size, but potentially large in terms of savings (e.g. reductions in the number of people claiming benefits), there may be a strong case for conducting an RCT.

Step 5: Assign each unit to one of the policy interventions, using a robust randomisation method

Random allocation of the units of study into policy intervention and control groups is the key step that makes the RCT superior to other types of policy evaluation: it enables us to be confident that the policy intervention group and control group are equivalent with respect to all key factors. In the context of education policy, for example, these might include socioeconomic status, gender, and previous educational attainment.

There are various ways that bias can creep in during the randomisation process, so it is important to ensure this step is done correctly from the outset, to avoid problems further down the line.

There is a lot of evidence that people who have a vested interest in a study may try to allocate people in a non-random manner, albeit unconsciously. For example, if a trial is allocating people to a "back-to-work" intervention on the basis of their National Insurance numbers, and odd numbers get the new intervention, then the person recruiting the participant may consciously or unconsciously exclude certain people with an odd NI number from the trial altogether, if they suspect they will not do well, in their desire to make the new intervention look good.

This introduces bias into the trial, and so the method of randomisation must be resistant to such interference. There are many independent organisations, such as clinical trials units, who can help to set up a secure randomisation service to avoid this problem of "poor allocation concealment". Typically this will involve a random number generator that determines which group a participant will be allocated to, and only after they have been formally recruited into the trial (for the reasons described above).

At the time of randomisation, if it is felt to be important, then steps can also be taken to ensure that the groups are evenly balanced with respect to various different characteristics: for example, to make sure that there is roughly the same age and sex distribution in each group. This is particularly important in smaller trials as thees have less power.

Box 15: Building in variations to enable testing

Testing involves comparing the effect of one intervention (e.g. possible new policy) against another (e.g. present policy). A sound test obviously requires that variations on the policy (e.g. new and present) can be delivered simultaneously. In some cases this is quite simple – some schools might continue to serve the usual school meals, while others could provide meals adhering to new nutritional standards, and the effect on classroom behaviour could be measured. In other cases, the systems in place may make it difficult to offer different policy variations at the same time.

For example, although a local authority may wish to test the effectiveness of simplifying a claim form, their letter systems may be outsourced, and/or incapable of printing more than one letter template. For this reason, we strongly suggest that when developing new systems or procuring new contracts from service providers, policy makers ensure that they will be able to deliver policy variations in the future. Although this may come at a slight upfront cost, the ability to test different versions of policies in the future is likely to more than justify this. With precisely this in mind, DWP legislation specifically allows the IT systems which deliver Universal Credit to include the facility to provide variations, to ensure that the department is capable of testing to find out what works, and adapting their services to reflect this.

Step 6: Introduce the policy interventions to the assigned groups

Once individuals, institutions or geographical areas have been randomly allocated to either a treatment group or a control, then it is time to introduce the policy intervention.

This might involve, for example, introducing a new type of education policy to a group of schools, while not making the corresponding changes elsewhere. When the Behavioural Insights Team ran a trial looking at whether text messages might improve people’s propensity to pay their Court fines, for example, individuals in the intervention groups received one of several different types of text message, whereas those in the control group received no text.

One important consideration at this stage is to have a system in place for monitoring the intervention, to ensure that it is being introduced in the way that was originally intended. In the text message example, for instance, it was useful to ensure that the right texts were going to the right people. The use of a process evaluation to monitor that the intervention is introduced as intended will ensure that results are as meaningful as possible and early hiccups can be rectified.

As with other steps, however, it will be important to ensure that it is possible to evaluate the trial in a way that reflects how it is likely to be rolled out if and when it is scaled up. For example, in the text message trial, it emerged that we did not always have the correct mobile numbers for everyone in the groups.

It would have been tempting to spend additional time and money to check and chase these additional telephone numbers, but it would not have reflected how the intervention might have been introduced if it were to be scaled up and would have therefore made the results appear more successful than they would be in “real life”.

Learn

Step 7: Measure the results and determine the impact of the policy interventions

Once the intervention has been introduced we need to measure outcomes. The timing and method of outcome assessment should have been decided before randomisation. It will depend upon how quickly we think the intervention will work, which will differ for each intervention.

A trial of different letters to people encouraging them to pay their fines may only need several weeks' follow-up, whilst a curriculum intervention may need a school term or even several years.

In addition to the main outcome, it may be useful to collect process measures. For example, in a study of differing probation services, one might collect data on referrals to different agencies to help explain the results. In this instance, a reduction in reoffending might be accompanied by a corresponding increase in referrals to anger management classes that might help explain the results. These secondary findings cannot be interpreted with the same certainty as the main results of the trial, but they can be used to develop new hypotheses for further trials (see Box 16).

In addition, many trials also involve the collection of qualitative data to help explain the findings, support future implementation, and act as a guide for further research or improving the intervention. This is not necessary, but if qualitative research is planned anyway, it is ideal to do it in relation to the same participants as those in the trial, since there may then be more information available.

Box 16. Smarter use of data

We are often interested in whether a policy intervention is broadly effective for a representative sample of the general population. In some cases, however, we might be interested to find out whether some groups (e.g. men and women, young and elderly people) respond differently to others. It is important to decide at the outset if we are interested in segmenting the sample in this way – if we do this after the data has been collected, sub-group analyses run a high risk of lacking both statistical power and validity. However, should a sub-group trend arise unexpectedly (e.g. men might seem to be more responsive than women to text message reminders to attend GP appointments), we might consider conducting a trial in the future to find out whether this is a robust result. It is usually worthwhile collecting additional data (e.g. age, gender) which will help you to segment your sample and inform future research.

Sometimes, an unanticipated trend may emerge from your trial data. For example, you might notice large fluctuations over time in the effectiveness of an incentive to take up loft insulation, and discover that this is relates to temperature variations. This trend might suggest that people are more receptive to messages about home insulation when the weather is cold. As it is an unplanned analysis, the results can't be considered definitive; however, no information should be wasted, and this result could be valuable to inform future research.

Adapt

Step 8: Adapt your policy intervention to reflect your findings

Implementing positive results about interventions is often easier than convincing people to stop policies that have been demonstrated to be ineffective. Any trial that is conducted, completed, and analysed, should be deemed successful. An RCT that shows no effect or a harmful effect from the new policy is just as valuable as one that shows a benefit.

The DWP trial of supporting people who were receiving sickness benefit was a "null" study in that it did not demonstrate effectiveness (see Box 2). However, if we can be confident that this was a fair test of whether the intervention works (which should be established before a trial commences), and that the sample size was large enough to detect any benefit of interest (which, again, should be established before commencing), then we have learnt useful information from this trial.

Where interventions have been shown to be ineffective, then "rational disinvestment" can be considered, and the money saved can be spent elsewhere, on interventions that are effective. Furthermore, such results should also act as catalysts to find other interventions which are effective: for example, other interventions to help people on sickness benefits.

When any RCT of a policy is completed it is good practice to publish the findings, with full information about the methods of the trial so that others can assess whether it was a "fair test" of the intervention. It is also important to include a full description of the intervention and the participants, so that others can implement the programme with confidence in other areas if they wish to.

A useful document that can guide the writing of the RCT report is the CONSORT statement[27], which is used in medical trials and also, increasingly, in non-medical trials. Following the CONSORT guidance will ensure the key parts of the trial and the interventions are sufficiently accurately described to allow reproduction of the trial or implementation of the intervention in a different area.

Ideally, the protocol of the trial should be published before the trial commences, so that people can offer criticisms or improvements before the trial is running. Publishing the protocol also makes it clear that the main outcome reported in the results definitely was the outcome that was chosen before the trial began.

Step 9: Return to Step 1 to continually improve your understanding of what works

Rather than seeing an RCT as a tool to evaluate a single programme at a given point in time, it is useful to think of RCTs as part of a continual process of policy innovation and improvement. Replication of the results of a trial is particularly important if the intervention is to be offered to a different population segment than that was involved in the original RCT. It is also useful to build on trial findings to identify new ways of improving outcomes.be particularly pertinent when RCTs are used to identify which aspects of a policy are having the greatest impact. In recent work with HMRC, for example, the Behavioural Insights Team has been attempting to understand which messages are most effective at helping people to comply with the tax system.

Several early lessons have been learnt about what works best – for example, keeping forms and letters as simple as possible and informing debtors that most others in their areas have already paid their tax.

However, rather than banking these lessons and assuming that perfection has been achieved, it is more useful to think of the potential for further refinement: are there, for example, other ways that we can find to simplify forms and make it easier for taxpayers to comply, or are there other messages that might resonate with different types of taxpayer?

The same type of thinking can apply to all areas of policy – from improving examination results to helping people into sustainable employment.

Continual improvement, in this sense, is the final, but arguably most important, aspect of the 'test, learn, adapt' methodology as it assumes that we never know as much as we could do about any given area of policy.

Box 17. Reducing patient mortality in nursing homes

Flu vaccinations are routinely offered to at risk groups, including the elderly, as the flu season approaches. In nursing homes however, the flu virus is likely to be introduced into the home via staff. An RCT was conducted in 2003 to determine whether the cost of a drive to vaccinate staff would a) increase staff vaccination rates, and b) have positive effects on patient health. Over 40 nursing homes were randomly allocated either to continue as usual (without a staff vaccination drive) or to put in place a campaign to raise staff awareness of flu vaccines and offer appointments for inoculation. Over two flu seasons, staff uptake of vaccines was significantly higher in nursing homes who instituted the staff flu campaign, perhaps unsurprisingly. Most importantly, the all-cause mortality of residents was also lower, with 5 fewer deaths for every 100 residents.[28] This research contributed to a national recommendation to vaccinate staff in care home settings, and is cited as part of the justification for continued recommendations to vaccinate healthcare workers internationally.

References

1. The first published RCT in medicine is credited to Sir A. Bradford Hill, an epidemiologist for England's Medical Research Council. The trial, published in the British Medical Journal in 1948, tested whether streptomycin is effective in treating tuberculosis.

2. Banerjee, A., Duflo, E. (2011). Poor economics: A radical rethinking of the way to fight global poverty PublicAffairs: New York.

3. Karlan, D. & Appel, J. (2011). More than good intentions: How a new economics is helping to solve global poverty. Dutton: New York.

4. Shepherd, J. (2007). The production and management of evidence for public service reform. Evidence and Policy, Policy Press Vol. 3 (2) pages 231-251

5. Department of Work and Pensions, Research Report 342, 2006. Impacts of the Job Retention and Rehabilitation Pilot. http://research.dwp.gov.uk/asd/asd5/rports2005-2006/rrep342.pdf

6. This is an interaction design, which allows the determination of the separate and combined effects of two interventions. Such designs are especially useful where questions exist about the additional effect of one/more features of a complex programme.

7. Department of Work and Pensions, Research Report 382, 2006. Jobseekers Allowance intervention pilots quantitative evaluation. http://research.dwp.gov.uk/asd/asd5/rports2005-2006/rrep382.pdf

8. Harford, T. (2011). Adapt: Why success always starts with failure. Little, Brown: London.

9. Taleb, N. N. (2007). The Black Swan: The impact of the highly improbable. Allen Lane: London.

10. Christensen, C. (2003). The Innovator's Dilemma: The revolutionary book that will change the way you do business. HarperBusiness: New York.

11. Luca, M. (2011). Reviews, reputation, and revenue: The case of Yelp.com. Harvard Business School Working Paper, No. 12-016.

12. Banerjee, A. V., Cole, S., Duflo, E. & Linden, L. (2007). Remedying education: Evidence from two randomised experiments in India. Quarterly Journal of Economics, MIT Press, vol. 122(3), pages 1235-1264

13. Davenport, T. H., & Harris, J. G. (2007). Competing on analytics: The new science of winning. Harvard Business School Press,

14. Delta Airlines Magazine (2007), 0915, 22.

15. Edwards,. P. et al. (2005). Final results of MRC CRASH, a randomised placebo-controlled trial of intravenous corticosteroid in adults with head injury – outcomes at 6 months. Lancet, 365, 1957-1959.

16. This can be estimated formally: by comparing the cost of a trial, for example, against an estimate of the money that would be wasted if the intervention was implemented but had no benefit.

17. Finckenauer J. O. (1982) Scared Straight and the Panacea Phenomenon. Englewood Cliffs, NJ: Prentice-Hall, 1982.

18. Petrosino, A., Turpin-Petrosino, C., & Buehler, J. (2003). Scared Straight and other juvenile awareness programs for preventing juvenile delinquency. Campbell Review Update I. The Campbell Collaboration Reviews of Intervention and Policy Evaluations (C2-RIPE). Philadelphia, Pennsylvania: Campbell Collaboration.

19. The Social Research Unit (2012). Youth justice: Cost and benefits. Investing in Children, 2.1 (April). Dartington: The Social Research Unit. Retrieved from http://www.dartington.org.uk/investinginchildren

20. Brooks, G., Burton, M., Cole, P., Miles, J., Torgerson, C. & Torgerson, D. (2008). Randomised controlled trial of incentives to improve attendance at adult literacy classes. Oxford Review of Education, 34(5), 493-504.

21. For a summary of the US research on the Family Nurse Partnership, see: MacMillan, H. L. et al. (2009). Interventions to prevent child maltreatment and associated impairment. Lancet, 363 (9659), 250-266.

22. http://www.campbellcollaboration.org/library.php

23. Final results yet to be published. For details on the study design, see: Pearson, D., Torgerson, D., McDougall, C., & Bowles, R. (2010). A parable of two agencies, one of which randomises. Annals of the American Academy of Political & Social Sciences, 628, 11-29.

24. Riggs, B. L., Hodgson, S. F., O'Fallon, W. M. (1990). Effect of fluoride treatment on fracture rate in postmenopausal women with osteoporosis. New England Journal of Medicine, 322, 802–809; Rothwell, P. M. (2005). External validity of randomised controlled trials:"To whom do the results of this trial apply?" Lancet, 365, 82-93

25. Cotterill, S. John, P., Liu, H., & Nomura, H. (2009). How to get those recycling boxes out: A randomised controlled trial of a door to door recycling service; John, P., Cotterill, S., Richardson, L., Moseley, A., Smith, G.,Stoker, G, & Wales, C. (2011). Nudge, nudge, think, think: Using experiments to change civic behaviour. London: Bloomsbury Academic.

26. Miguel, E. & Kremer, M. (2004). Worms: Identifying impacts on education and health in the presence of treatment externalities. Econometrica, 72, 159-217.

27. http://www.consort-statement.org/consort-statement

28. Hayward, A. et al. (2006). Effectiveness of influenza vaccine programme for care home staff to prevent death, morbidity and health service use among residents; cluster randomised control trial. British Medical Journal, 333 (7581), 1241-1247.

CabinetOffice
Behavioural Insights Team

Published by the Cabinet Office Behavioural Insights Team

Publication date: June 2012

This document can also be viewed on our website at www.cabinetoffice.gov.uk