Alliance for Useful Evidence

The Experimenter's Inventory

A catalogue of experiments for decision-makers and professionals

Anna Hopkins, Jonathan Breckon and James Lawrence

January 2020

Funded by:

nesta

About the Alliance for Useful Evidence

The Alliance for Useful Evidence is a network, hosted by Nesta, that champions the smarter use of evidence in social policy and practice. We do this through advocacy, convening events, sharing ideas and resources, and supporting individuals and organisations through advice and training. We promote our work through our network of more than 4,800 individuals from across government, universities, charities, businesses, and local authorities in the UK and internationally. Anyone can join the Alliance network at no cost.

To sign up please visit:
www.alliance4usefulevidence.org/join

About the authors

Anna Hopkins is a Researcher at the Alliance for Useful Evidence at Nesta.
Jonathan Breckon is Director of the Alliance for Useful Evidence at Nesta.
James Lawrence is the Head of Quantitative Research at the Behavioural Insights Team, which is part of the Alliance for Useful Evidence network.

How to cite this inventory

Hopkins, A., Breckon, J. and Lawrence, J. (2020) *The Experimenter's Inventory: A catalogue of experiments for decision-makers and professionals*. Alliance for Useful Evidence, Nesta, London.

Acknowledgements

We are especially grateful to **Eszter Czibor** (Innovation Growth Lab) and **Guillermo Rodriguez-Guzmán** (Education Endowment Foundation) for their comments on this inventory. This report draws on discussions with Nesta's experimentation working group: with thanks to **Albert Bravo-Biosca**, **Jesper Christiansen**, **Teo Firpo**, **Bas Leurs** and **James Oriel** for their time, input and advice. Thanks also for illuminating conversations on the ideas in this report from **Rob Ashelford** (Y-Lab, Nesta), **Mikko Annala** (Demos Helsinki), **Jon Baron** (Arnold Ventures), **Pierre-Olivier Bédard** (Treasury Board of Canada), **Chris Bonell** (London School of Hygiene and Tropical Medicine), **Jon Brown** (NSPCC), **James Cairns** (Center on the Developing Child at Harvard), **Nick Chesterley** (Treasury Board of Canada), **Geoffrey M. Curran** (University of Arkansas for Medical Sciences), **Carrie Deacon** (Nesta), **Brad Dudding** (Centre for Employment Opportunities), **Triin Edovald** (Education Endowment Foundation), **Dan Farag** (Nesta), **Jen Gold** (Cabinet Office), **Beatriz Hasbun** (Laboratorio de Gobierno), **Tim Hobbs** (Dartington Service Design Lab), **Myra Latendresse-Drapeau** (Employment and Social Development Canada), **Helen Mthiyane** (Alliance for Useful Evidence), **Geoff Mulgan** (Nesta), **Jenny North** (Dartington Service Design Lab), **Sara Peters** (Project Evident) and **Julie Reed** (Imperial College London) and **Alex Sutherland** (Behavioural Insights Team).

If you'd like this publication in an alternative format such as Braille or large print, please contact us at: **information@nesta.org.uk**

Alliance for Useful Evidence

The Experimenter's Inventory

A catalogue of experiments for decision-makers and professionals

Alliance for Useful Evidence

Summary

Section 1: Introduction

Experiments have a rich history. From some of the earliest scientific breakthroughs, through to the new institutions of policy evaluation, experiments have played a central role in social and technological change. **Section 1** of this report explores this history, and the role that experiments can play in public problem-solving today.

Section 2: What is an experiment?

Section 2 outlines three basic experimental designs.

- **Randomised experiments** are often considered the 'gold standard' of evaluation. New designs are making them easier to run, more sensitive to context, and better at helping us tackle complex problems.
- **Non-randomised and quasi-experimental designs** are a growing range of approaches that use inventive designs, 'big data' and statistical methods to put policy to the test.
- **Pre-experimental** approaches are being used to explore and develop novel ideas – and find out if new solutions are feasible in practice.

Section 3: An inventory of experimental approaches

Our inventory provides a framework for thinking about the choices available to an agency, ministry or organisation that wants to experiment. **Section 3** describes each design in plain language, highlights great examples and useful applications, as well as the limitations of different designs.

Randomised experiments	1. The basic RCT 2. Multi-arm trial 3. Nimble RCT 4. A/B test 5. Cluster randomised trial 6. Stepped wedge and wait-list designs 7. Cross-over design 8. Multi-site trial 9. Realist trial 10. Hybrid trial 11. Adaptive trial
Non-randomised and quasi-experimental designs	12. Regression discontinuity design (RDD) 13. Matching 14. Difference-in-difference (DiD) 15. Synthetic control
Pre-experiments	16. Pre-post test 17. Rapid cycle testing 18. Prototyping

Section 4: Useful resources

This inventory serves as an introduction to a wide range of experimental approaches. For those who want to learn more, or get support doing an experiment, we provide a list of further resources.

The annexes

Experimentation jargon buster

Annex A clarifies the more technical language used in the report. Words that feature in this glossary are written in *italics*.

Some common criticisms of RCTs

Annex B covers some common criticism of RCTs – and responses.

What is experimentation in policy or practice? And why is an experiment worth investing in? What can an experiment tell you? And when or where might it fall short?

Answers to these questions are often complicated and unclear. This inventory provides an antidote: a catalogue of experiments of different shapes and sizes, and simple advice on the pros and cons of different designs.

What's in this inventory?

This inventory is about how you can use experiments to solve public and social problems. It aims to provide a framework for thinking about the choices available to a government, funder or delivery organisation that wants to experiment more effectively. We aim to simplify jargon and do some myth-busting on common misperceptions.

There are other guides on specific areas of experimentation – such as on randomised controlled trials – including many specialist technical textbooks. This is not a technical manual or guide about how to run experiments. Rather, this inventory is useful for anybody wanting a jargon-free overview of the types and uses of experiments. It is unique in its breadth – covering the whole landscape of social and policy experimentation, including prototyping, rapid cycle testing, quasi-experimental designs, and a range of different types of randomised trials. Experimentation can be a confusing landscape – and there are competing definitions about what constitutes an experiment among researchers, innovators and evaluation practitioners. We take a pragmatic approach, including different designs that are useful for public problem-solving, under our experimental umbrella.[i] We cover ways of experimenting that are both qualitative and quantitative, and highlight what we can learn from different approaches.

At **Nesta**, we are interested in both the theory and practice of experiments. We have run several of our own, as well as working to grow their use in new policy areas and exploring novel platforms like experimental testbeds and funds.[1] In 2009, we ran one of the earliest randomised controlled trials on a business support scheme, a kind of innovation voucher called 'Creative Credits'. It has spawned many similar initiatives around the world. A few years later, Nesta launched the **Innovation Growth Lab** (IGL), the biggest global partnership supporting and running randomised controlled trials on economic, innovation and industrial policy. IGL has supported the investment of more than $2.8 million in experiments with organisations in 26 countries. Nesta's Innovation Skills team provides training for public servants internationally, with a focus on problem-solving, and has published guides that aim to make some of the tools covered in this report, like prototyping, more do-able for practitioners in government, charities and foundations.

The **Alliance for Useful Evidence** is a UK-wide network hosted by Nesta, dedicated to championing the use of evidence in social policy. In 2015, we published *Better Public Services Through Experimental Government*, which made the case that government must rigorously and systematically put policy to the test – or risk stagnation.[2] With Nesta, we have worked to promote and embed experimentation within policy and practice, launching the UK's What Works Network. In our work with partners, advising governments and foundations on how to test new ideas, many asked us for a comprehensive guide that brings different experimental approaches together.

This inventory draws on lessons from a broad range of experimenters – professionals in charities, change-makers in frontline services, researchers in academic institutes, and decision-makers in government – who are pioneering the use of experiments to create better solutions to the challenges of the twenty-first century.

Our focus is primarily on social policy – but we also share lessons from medicine, public health, business and international development. Some of our discussion focuses on our experiences here in the UK. But, we have aimed to look internationally, to learn from experimenters around the world.

Section 1: Introduction

Experimentation has a rich history – a history that, in many ways, has shaped the trajectory of science and society. From Galileo's use of rolling balls to explore the laws of physics, through to the thousands of trial-and-error experiments conducted by Thomas Edison to create the first lightbulb, trying things out in practice has been a cornerstone of scientific and technological discovery.[3]

As science and social science matured, experiments came to play a significant role in underpinning the philosophy of science, which investigates the best ways to develop valuable new knowledge. Karl Popper argued in his 1959 *The Logic of Scientific Discovery* that to be scientific, theories must be falsifiable – that is, science shouldn't be murky, confusing or circular. Popper argued that to do science we must put bold and imaginative ideas to the test, improving them ruthlessly through empirical investigation.[4] To develop robust ideas we must constantly try to prove ourselves wrong. But experiments aren't only about distinguishing fact from fiction – they are tools for exploring the unknown. As social psychologist Karl Lewin wrote of his early experiments, *"if you want truly to understand something, try to change it."*[5]

Experiments have entered the mainstream in science, medicine and some areas of international development. In health, innovation and experimentation now go hand in hand – from trials that develop effective treatments, through to tests of new approaches at the frontline of services. In 2019, three pioneers of randomised experiments – Esther Duflo, Abhijit Banerjee and Michael Kremer – won the Nobel Memorial Prize in Economics. They set up the Abdul Latif Jameel Poverty Action Lab (or JPAL), an organisation which has run more than 1,000 randomised controlled trials to understand how to reduce poverty and has championed the use of the method internationally.

Business has caught on too. Among the largest financial institutions, retailers and restaurants in the US, at least a third are running randomised experiments.[6] A/B testing is now the standard (although rarely advertised) tool used by Silicon Valley to improve its online products.[7] Companies like Amazon and Google run tens of thousands of experiments a year.[8]

Experimentation is however still rare in government decision-making in the organisations, services and areas of policy that tackle the most pressing problems that face our societies. Whole swathes of domestic public policy – like skills and employment, policing and crime, and almost all science and innovation funding – remain relatively untouched by experimentation. This inventory sets out the tools we have available, and how experimentation can become mainstream in how we solve public problems.

Why experiment?

What makes experiments useful

Experiments share common characteristics that make them especially valuable for science and problem-solving. They break down big issues into smaller questions that can be more manageably investigated, in a way that is structured and transparent.[9] They do this by establishing a clear idea, carefully defined, which can be tested or trialled. Experiments have an unambiguous structure: fixed timelines, limits and checkpoints are established, at which results are assessed and decisions made. These are agreed at the start and can't be changed at whim or manipulated to suit the experimenter. This is why experiments have for centuries been so highly prized: they set out a process for systematic and transparent learning. They aren't PR stunts, which guarantee positive outcomes to be used for show, but generate practical new information to help shape wise choices. These attributes have long been recognised – and provide an unbiased foundation for policy and debate. Professor Alvin Roth, Nobel Laureate and former President of the American Economic Association, has written that experiments can 'speak to theorists', 'search for facts' and 'whisper in the ear of princes'.[10]

Learning from failure – a more humble approach

Experimentation allows us to learn positively from failure. Amy Edmondson at the Harvard Business School argues that we should make a distinction between good and bad failure. Good failure is part of the unavoidable process of learning and exploring. Bad failure, Edmondson tells us, is preventable failure, that doesn't result in new learning.[11] This distinction contains useful lessons for teams faced with tough choices and uncertainties, and requires a more reasonable attitude to risk from decision-makers and leaders.

An experimental mindset advocates humility: we must take a more humble approach to learning empirically about the world, rather than assuming that we know all the answers already.[12] British political scientist Gerry Stoker has argued that experiments are useful because they help break up complex issues into tractable questions that can be explored and investigated.[13] Probing and experimenting may be the best way to learn in complex, uncertain systems – although we still have much to learn about how to do this well.[14]

Finding out what doesn't work, as well as what does

Crucially, experimentation means being transparent about what doesn't work, as well as what does. There is no avoiding the challenges this brings; and it may mean taking a more honest approach – what Jen Gold, Head of the Cabinet Office's What Works Team, has called 'open-by-default'[15] – recognising in public that you don't have all the answers, and need to test out your ideas. Dr Gold points to the Canadian government's Experimentation Works programme (Box 1) and the US' Office of Evaluation Sciences (OES) as exemplars of this attitude.

Box 1: Experimentation Works: 'open-by-default' learning in Canada

"How do we put in place systems that help everybody experiment?" Canada's Treasury Board Secretariat houses a growing Innovation and Experimentation Team, with a mandate to support experimentation across the Government of Canada. Director Nicholas Chesterley, along with another of Canada's central agencies, the Privy Council Office, is leading Canada's strategy on experimental policy – and searching for a more open approach.[16]

Experimentation Works (EW) aims to build public servants' capacity in experimentation through a learning-by-doing model. Cohorts of public servants work on experimental projects, and EW shares practical learning and examples, and ensures open access to learning materials, progress updates and results. It works by connecting project teams with each other and with experts, where groups learn together and share experiences.

This open approach is helping departments and agencies identify gaps in expertise, lack of capacity for partnerships, and where and why teams struggle to take successful experiments to scale. It's helping them learn more quickly about the nuts and bolts of what experimenting means in practice. One example is Health Canada's PRODigy experiment team, who are running A/B tests on their website to see if changes in layout and language can improve incident reporting rates.[17] They have come across some unexpected bumps in the road, and lots of what they have learnt is practical – like how best to work with colleagues in different areas of the department, and how to know when you need expert advice. These insights are now captured and available publicly on the EW blog.[18]

Holding decision-making to account

The taste for 'big bang' policymaking that often seems ubiquitous in government – the big policy announcement, the unstoppable rollout – is being challenged. Universal Credit seems a prime example of this political grandstanding in action – a benefits system overhaul ostensibly piloted by the UK Government, but hailed as a flagship policy, a 'once-in-many-generations reform', before it was trialled.[19] It now threatens to fail its proponents spectacularly.[20] The risks of over-investing in untested policy are obvious. The political, reputational and economic costs are high, as are the chances of preventable failure.

Experiments teach us to take a different attitude, summed up in the 'test, learn, adapt' mantra coined by the Behavioural Insights Team. Here, failure can be embraced as intelligent, structured and mitigated – although not always without discomfort.[21] This approach aims to build an effective product and service from repeated trialling, testing and evaluation. Starting small can allow organisations and departments to give innovations a better chance of success and provide a more cost-effective way of developing new ideas, trying them out in practice to see if they are worth investing in. It also allows us to learn more about innovations, tweaking and improving them before they are rolled out to communities.

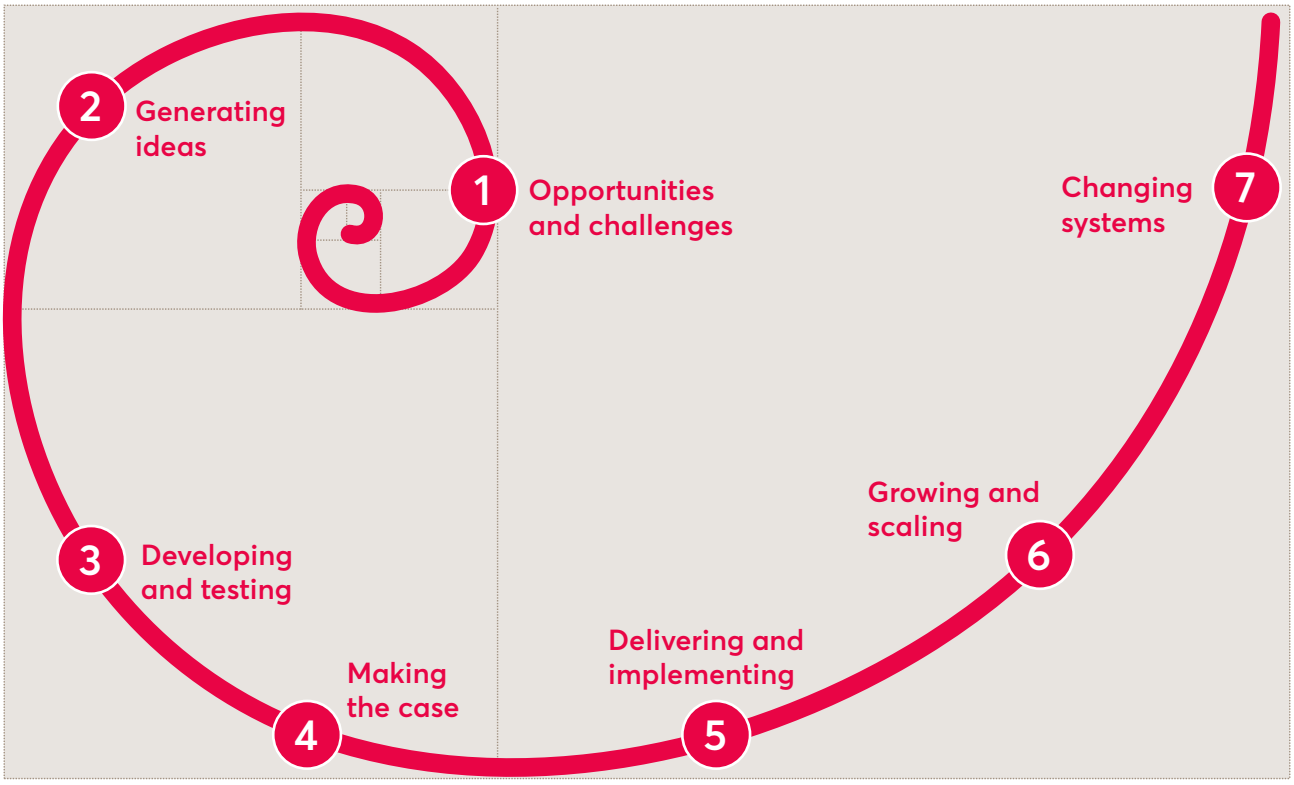
Examples of this approach are now reaching policy, like the trial of Magic Breakfast run by the Education Endowment Foundation, the Department for Education and the Behavioural Insights Team, which grew from a test of breakfast clubs in 106 schools to an investment of £26 million in morning clubs in more than 1,770 schools, focused on the most disadvantaged areas.[22] But these examples are rare.

What's in our toolbox?

This inventory shows that we have more tools to hand than ever. But do we always know which to use, and when? Nesta's innovation spiral maps the process of innovation. Starting with identifying opportunities and generating new ideas, it shows how innovations can be grown through a testing and learning approach. Experiments don't come out of nowhere, and are only relevant at certain key moments in the lifespan of an idea. To build solutions worth testing, we must draw on past evidence and conduct new research. This groundwork is what helps us come up with a promising idea, as well as helping us set our *hypothesis* – the research question that an experiment aims to answer. Experimentation becomes most important when we already have a clearly defined idea and want to develop it, try it out in practice, or see whether it works.

Some of the pre-experimental methods we cover in this report – such as prototyping – are a good fit for the earlier stages on the innovation spiral (1 to 4); they help us explore new ideas and come up with hypotheses. Later in a project, when a policy idea or service is being trialled or grown, quasi-experimental or randomised experiments might be more appropriate. Experiments are also crucial at the point of growing or scaling ideas (6 on the spiral). They tell us what worked and what didn't, where something can be improved, and how money should be invested. Can something that seemed to work in Manchester also work in Aberdeen, Cardiff or Belfast? We should aim to replicate experiments – run them in different times and places – to learn about how to grow, scale and adapt our ideas.

Figure 1: Nesta's Innovation Spiral



Source: Nesta

The important principle is 'horses for courses' – matching your challenge to the right type of experiment. To help you here, you can get a feel for the benefits and downsides of each approach in **Table 1**, which summarises all 18 experimental approaches in our inventory.

The world of experiments can be tough to navigate. Anyone who has been involved in policy development or evaluation will know it can be divisive territory. Randomised controlled trials (or RCTs), the so-called 'gold standard' of impact evaluation, are like the British sandwich spread Marmite – people tend to either love them or hate them.[ii] But polarised opinions do not often help us make appropriate choices.[23] A wise approach is to be open to the breadth of methods, but to think carefully about which ones are best suited to getting the answers we need. Few researchers now would advocate the same approach to running experimental evaluations as the 'randomistas' of the 1960s.[iii] We should be clear about the value of randomised experimental designs but also be aware of when more flexible approaches may be advantageous.

A very brief history of experiments in decision-making

From 400 years ago, thinkers like Sir Francis Bacon, in his book *New Atlantis*, began to apply these ideas to the principles of government. He argued that we must challenge the common-sense assumptions we gain from everyday experience by testing and learning, and set out a radical vision for an experimental state.[24] But it was centuries until the ideas and principles of science were applied to making policy and creating social change.

Governments have however embarked on experiments of many kinds; in a sense, most governments are constantly 'experimenting', trying out new things, but without learning from them. Many of the catastrophic consequences brought about by irresponsibly 'trying stuff out' on whole populations are catalogued in the book *The Blunders of our Governments*, which covers many of the huge UK policy failures resulting from untested ideas, from the Poll Tax to Individual Learning Accounts.[25] Despite the promise of experiments to help us be more effective and efficient, many organisations and governments have failed to embrace a mindset of testing and learning.

In fact, it was in agriculture that the experimental method was first systematically applied to planning. At the Rothamsted Experimental Station in Hertfordshire, one of the oldest agricultural research centres in the world, scientists Sir John Bennet Lawes and Joseph Henry Gilbert began the first large field experiment to study how best to grow healthy crops. Seven of these classic experiments continue today.[26] It was in agriculture too that the principles of the experimental method – in the formal terms of social science – were first established, in Ronald Fisher's foundational work *The Design of Experiments*.[27]

The history of experimentation does, of course, look very different across the world. China, for example, has run many large-scale experiments on major economic policies. Professor Sebastian Heilmann, Director of the Mercator Institute for China Studies, has pointed out that many of the Communist party-state regulations throughout the 1980s and 1990s were 'experimental'. In 1988, 'experimental zones' were set up, like the Shenzhen Special Economic Zone near Hong Kong, which took different approaches to encouraging growth.[28] But these weren't experiments in the scientific sense of the word – and there was little public evidence or data to share from these policy changes. Practices such as these are so common that the word 'experimental' has come to mean 'innovative' or 'radical', rather than simply 'untested'.

More cautious, structured experiments first found their way into policymaking in 1930s America, under the leadership of Franklin D. Roosevelt who called for 'bold, persistent experimentation' with the social programmes of the New Deal.[29] This first era inspired a generation of skilled policy experimenters in the 1960s, 'randomistas' like Judith Gueron, former President of the US MDRC (founded in 1974 as the Manpower Demonstration Research Corporation), who fought for reliable evidence to become the basis of government decision-making and underpin the responsible stewardship of public funds.[30] In 1971, RAND began a randomised experiment about health insurance, and who should pay for it, in a trial that spanned more than a decade, and remains one of the largest experiments ever conducted.[31] At the same time, US social scientists like Donald Campbell argued for experiments to be part of a new approach to government. Campbell's vision for an 'experimenting society' inspired the Campbell Collaboration, which now works to grow the use of experimental results in social policy. But the vision of Campbell and others was not realised. In fact, this first phase of experimentation in government met challenges. Some researchers over-promised on what they could deliver, and policymakers became frustrated with waiting for experiments to yield results.[32]

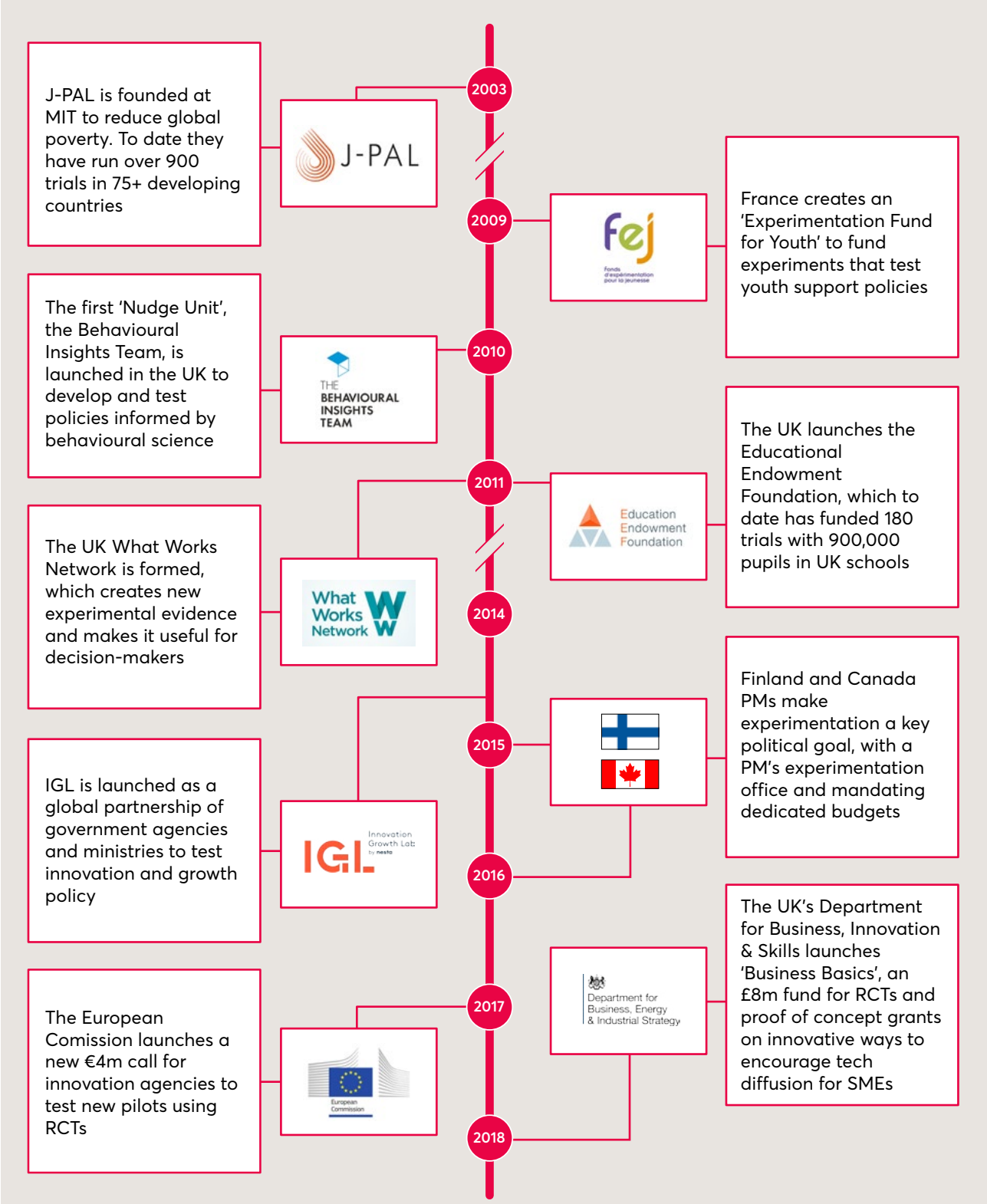
After this second wave, experimentation in policy went quiet. Then, in the 1990s, we began to see randomised controlled trials (or RCTs) being used in international development. Organisations like Innovations for Poverty Action and the aforementioned Abdul Latif Jameel Poverty Action Lab (J-PAL) were founded in the early 2000s to answer critical questions about how best to reduce poverty. In 2004, the World Bank set up the Development Impact Evaluation programme, DIME; and in 2006, the Washington-based think tank the Centre for Global Development (CGD) published an influential report, *"berating the development community for spending billions of dollars on programmes for which there was no evidence."*[33] The infrastructure they put in place has enabled many other development researchers to conduct randomised experiments.

Mexico provided an early example of how experimentation can be done on an ambitious national scale. The social assistance programme Progresa (now called Prospera), which reached more than five million families between 1997 and 2005, took a pragmatic approach to trying out different approaches on the ground and rigorously evaluating them. Providing conditional cash transfers was found to have strong positive effects on education, health, nutrition and poverty. This evidence has helped it survive changes of government – and rebranding – and led to its adoption in other Latin American countries.[34] It's unclear if this political support will survive – Mexico's 'model for the world' is now under threat of abolition, reminding us that even strong evidence can be ignored.[35] Here in the UK, political support for experimentation has been slow to develop. After a faltering start, there was some take-up in the early 2000s, particularly after the 2003 publication of HM Treasury's *Magenta Book*, a guide on government evaluation.[iv]

The growing adoption of experimentation in the 21st century

Over the past two decades, we have seen experimentation spread and put down roots, in many countries across the world. In the UK, the What Works Centres have tested how best to deliver value for citizens through effective early years, education, local economic and social care policy. Experiments are also now more widely accepted by the public – the Education Endowment Foundation (EEF) has supported more than 180 randomised trials in over half of English schools.[36]

Figure 2: The growing adoption of experimentation



Source: Innovation Growth Lab, Nesta

Experimental teams dedicated to testing and evaluating new approaches to public service delivery now sit in governments across the world including Canada, the USA, Australia, Colombia and the UAE.[37] In Finland, experimentation has been elevated to official government policy, through a team and strategic mission housed at the Prime Minister's Office.[38] In France, the public policy laboratory Fonds d'Expérimentation pour la Jeunesse has funded experimental youth programmes and evaluations since 2009 – including the 'Garantie Jeune' or Youth Guarantee, which has extended unemployment support to young adults.[39] A trial helped the French government understand how best to support vulnerable young people. An increase in financial support meant that after 14 months 40.4 per cent of beneficiaries were in employment, compared with 34.1 per cent of the control group.[40] And in 2015, the Canadian Prime Minister issued a breakthrough mandate letter requesting that departments devote, *"a fixed percentage of program funds to experimenting with new approaches."*[41] These teams draw on different methods, from randomised trials to rapid tests, with some focused on policy evaluation and others, like Finland's citizen-centred model, aimed at building a culture of innovation and participation from the grass-roots.[42] Nesta's **Innovation Growth Lab** (**Box 2**) is working with governments worldwide to bring randomised experiments to an area of policy where billions of pounds are spent each year, but little evidence exists – innovation and growth.

Box 2: Testing innovation and growth policy

'Everyone agrees that innovation is vital to accelerate productivity growth and solve the big challenges of our era. The problem is that we don't really know what works and what doesn't to increase innovation.'[43]

Albert Bravo-Biosca & Lou-Davina Stouffs, IGL

European governments spend about €150 billion every year supporting entrepreneurs and businesses to innovate and grow. But very little is known about what's effective to support innovation, entrepreneurship and growth.

In 2014, Nesta launched the Innovation Growth Lab (IGL), a global collaboration that works to systematically develop and test new approaches to innovation and growth policy. They aim to find out what works and what doesn't (and when), learning from successful experiences in other fields, such as development economics, health and education.

IGL emphasises the importance of informed decision-making. For example, they are continuing to investigate the use of innovation vouchers, a popular policy among innovation agencies to connect small and medium-sized enterprises (SMEs) to expert knowledge providers. There is still little evidence on whether they achieve their ultimate goal: spurring firms to establish long-term relationships which help them innovate.[44] IGL has been supporting one of its partners, Innovate UK, to evaluate an Innovation Vouchers programme that has now been closed.[45] Based on the results, Innovate UK will be able to move forward with all the facts at hand and decide whether this is a programme worth repeating.

The use of experiments to guide effective social action is being championed, and re-invented, outside of government too. Organisations like Giving Evidence have advocated for charities – and donors – to 'get smarter' in how they use evidence to achieve their goals.[46] Charities in the UK have started to embrace testing and evaluation, with pioneers like the NSPCC, The Money Charity and Chance UK using experiments to improve their work with children and young people. NSPCC has conducted the largest multi-site randomised trial of a sexual-abuse therapy programme in the world and is exploring rapid ways of evaluating community-led initiatives.[v]

It's not only the classic social scientific experiment that has entered the toolbox of the policymaker. Prototyping is an approach traditionally used by engineers, designers and web developers, and is now part of the landscape of policy and service design and decision-making. A famous example of the method is provided by the Dyson vacuum cleaner. First mocked-up from cardboard with a vacuum pump cannibalised from another manufacturer's product, it went through more than 5,000 iterations before going to market. The idea of using design methods such as this in policymaking is a relatively recent one, encouraged by an agenda-setting book published in 2014 by Christian Bason.[47] Christian is the former Director of MindLab, Denmark's flagship government innovation lab (now reinvented as the Disruption Task Force). These tools are central to the work of innovation labs, often in-house teams who work with governments to help them adopt an approach that front-loads risk, and embraces 'good failures', by trying out ideas early. These Labs now exist around the world. While they aren't always focused on research and evidence, they have drawn attention to the need to foster the 'mindset', as well as methods, of experimenting.[48] One example is Nesta's **States of Change** collective, which runs international learning programmes to embed experimental thinking in the day-to-day work of government.[49]

Another set of approaches central to learning about policy are quasi-experimental designs (QEDs), sometimes called 'queasy' experiments – after the sense of unease they can instil in purist researchers who prefer randomised designs.[50] These methods have deep roots in social science and statistics, providing ways to test and learn about policy ideas that are already in train or in areas of decision-making where randomised approaches may not be possible. Donald T Campbell, one of the originators of the movement for experimental government we mentioned earlier, was an advocate of these designs – and despite a rather unfashionable status among policy wonks and evaluators, non-randomised approaches have demonstrated their value and utility in many areas of policy. These designs also allow us to exploit unplanned opportunities for learning about social change, in what are termed '*natural experiments*'. These are tests that explore a policy or practice change resulting from a 'natural' divergence or external event, like devolution in the UK. The impact of different policies can be compared, such as free tuition fees in Scotland, free prescription charges in Wales, or the introduction of a congestion charge in London.

The benefits of agility and responsiveness in experimental trials are also championed by the 'nudge *units*' which have taken Behavioural Science, a discipline that studies human motivation and action, out of the lab and into the real world. These teams – from Peru's MineduLAB to Australia's Behavioural Economics Team (BETA) – have run randomised experiments to improve citizens' experience of services and are now helping to make the back offices of public sector organisations more intelligent.[51]

In healthcare, where the randomised trial has a history dating back to the 1747 experiment by ship's surgeon James Lind,[52] there is impressive innovation taking place in how experiments happen and a raft of new designs being used. Many aim to tackle more pragmatically the challenges of learning about complex public problems. Some are designed to help us gain insights across places and populations, through randomised experiments that prioritise building theory and gaining insight across diverse locations (see Sections **3.8** and **3.9**). Others value local knowledge, situated in context, that's responsive to the questions and concerns of patients and professionals. This work points to the limitations of linear models of cause-and-effect, arguing that we need a more nuanced understanding of how social change happens, using a broad spectrum of methods to design, implement, and evaluate innovations.[53] Some of these experiments call on a different ethos of scientific enquiry – emphasising the need to work with (rather than against) complexity, and learn incrementally through trial and adaptation. They also point to the need for greater efficiency in all aspects of public services. Dr Don Berwick, the founding President of the Institute for Healthcare Improvement (IHI) in the US, has argued for a Science of Improvement that aims to raise the quality of whole systems of care.[54] Like the work of innovation labs, these principles emphasise the value of agile project management, and iterative trialling and assessment, to improving organisations and achieving impact.

Finding better ways to learn empirically about making change has meant breaking new ground for applied research and evaluation. In the US, philanthropic donors like The Edna McConnell Clark Foundation are investing in new approaches to experimentation and evidence, grown out of the challenges of creating change through cross-sector partnerships.[55] In Canada, not-for-profits and social enterprises are behind more dynamic approaches to social innovation and R&D (research & development). A 2016 Canadian report called for a new approach to creating social impact through *"[a] combination of competency, culture, and craft that is intentionally applied to continuously learn, evaluate, refine and conduct practical experiments in order to enhance social wellbeing."*[56]

There is much to learn from developments over the past two decades. This inventory advocates openness to the breadth of experimental tools we have at our disposal – and provides a guide on how to make informed choices about which tools to use when, for organisations who want to make a difference on the issues that matter.

Section 2: What is an experiment?

We start with this simple definition from the Collins English Dictionary:

> "An experiment is the trying out of a new idea or method in order to see what it is like and what effects it has."[57]

As a test or investigation, an experiment has a structure – a systematic process that enables us to learn. And, experiments generate evidence – new information that can be used in decision-making. They are a hands-on learning strategy, that make an organised intervention or change in the world and investigate what happens as a result. Experiments usually involve some kind of control – that is, they hold some factors constant, while changing others, to get a clearer picture of what happens when a change is made.[58] They let us try out ideas in practice, refining and improving them over time.

Among experimenters with different types of expertise – researchers, innovators or evaluators – there are competing definitions about what constitutes an experiment. Some would argue that only a randomised trial counts as a 'true' experiment, while others would advocate for a looser definition. We take a pragmatic approach, including different designs under our experimental umbrella,[59] but are clear about what we can (and can't) learn from them. When we talk about experiments in this inventory, we do not mean experimental 'freewheeling' or 'just trying stuff out', but rather systematic learning strategies that make us more effective.[60]

Three basic experimental approaches

Three designs underpin all the experimental approaches in this handbook, with the exception of some 'quick and dirty' forms of prototyping. There are important pros and cons to each design, and caveats to what we can learn from them. In their own way, each of these approaches aim to create what's called the *counterfactual*, an estimation of what would have happened if the innovation hadn't taken place. The *counterfactual* asks, 'what would have happened to those people who received this innovation, if the innovation had not taken place?'. This is the fundamental problem of causality, and making what researchers call *causal inference*. It's an issue that randomised experiments address through the use of a control group; quasi-experimental designs using statistical methods; and pre-experimental approaches by using an individual's past outcomes as their *counterfactual*. Each of these ways of investigating the *counterfactual* – 'what would have happened?' – has consequences and allows us to learn about cause-and-effect with different levels of reliability.[61] A breakdown of all the technical terms we use in the report can be found in Annex A.

Randomised experiments

Randomised experiments aim to test a policy idea or innovation by investigating what difference it has made for the people it is aiming to help. They do this by using a control group; this means that some people receive an innovation, while others don't. By comparing results for the groups who receive an innovation and those that didn't, evaluators and decision-makers can get a clear sense of what the impact of a project or policy has been.

Unlike other kinds of experiment, randomised experiments allocate the control and experiment groups by chance; and this is where the strength of their design lies, in its ability to reduce what is technically termed *selection bias*. *Selection bias* results from the experimental and control group being fundamentally different in some (often unobservable) way – and will skew results. When a sample is large enough, random allocation will even out differences between the control and experiment groups, to create a fair comparison. This makes randomised experiments particularly valuable for answering questions about cause-and-effect: because randomisation creates groups that are comparable before the innovation, any group-level differences we observe afterwards can reliably be attributed to the innovation. For this reason, randomised experiments are often described as the 'gold standard': their design, when conducted well, creates the fairest possible comparison.[62] They allow us to make strong connections between a cause (the innovation we are testing) and an effect (the outcome we are trying to change). As **Figure 3** shows, randomised experiments have the greatest 'causal power': this means that, in general, we can talk about their results with the most certainty.

Non-randomised and quasi-experimental designs

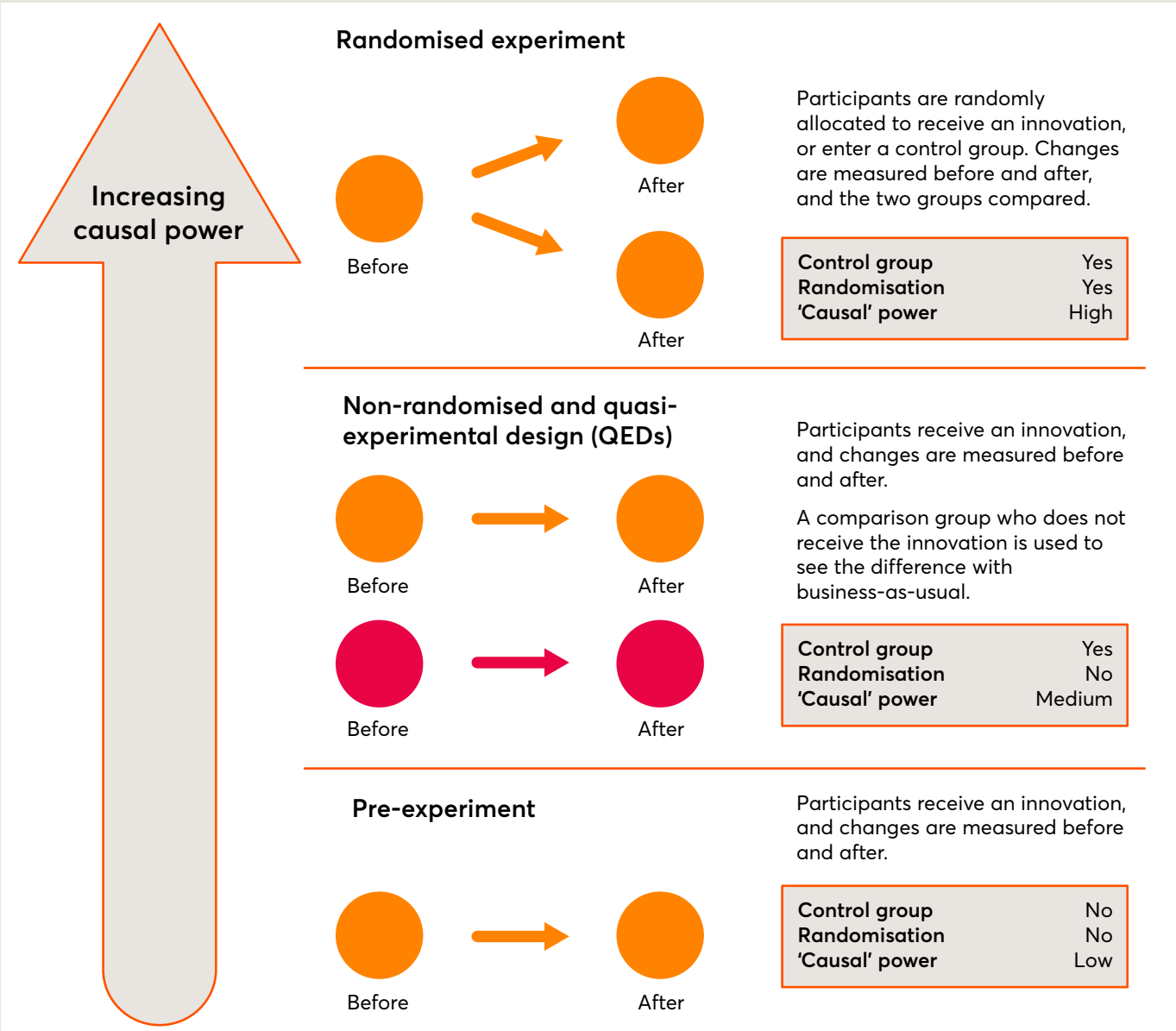
Quasi-experimental designs (QEDs) use statistical methods to create a comparison group, allowing us to learn about how an innovation works and what impact it has had. Most textbooks on QEDs are jargon-heavy and highly technical, which doesn't tend to make them appealing to a non-expert audience. We aim to clarify here. In essence, QEDs use various techniques to create a comparison group that is as similar as possible to the group receiving an innovation. This comparison then helps us investigate whether or not an innovation is really making a positive difference.

We include in this category some kinds of experiments that don't fall under the traditional banner of QEDs (researchers can be sticklers for detail), which we call 'non-randomised experiments'. Regardless, all of these designs aim to create a comparison group that's similar to the experimental group in terms of all of the characteristics we know about that might be relevant to the test – like socio-economic background, voting preferences or age. Some designs will also aim to create groups that are similar in terms of the unobservable characteristics, things that we don't know about, too. Like all research methods, these designs must be done well to yield useful results. A risk with QEDs is that they don't succeed in creating a similar enough comparison group. If this happens, it can introduce *bias* into the experiment, which may distort the results.

Pre-experiments

Sometimes called a 'within-subject design', pre-experiments compare one group of participants before and after an intervention, to see what's changed.[63] In his book on digital social science Professor Matthew Salganik terms these experiments 'perturb and observe' – they make a change in the world, and measure what happens, with a single group. By doing a before and after comparison, these designs use individuals' past (or 'before') outcomes, to estimate the *counterfactual*. 'Perturb and observe' experiments are useful when we are aiming to probe and discover, to shape hypotheses that can be tested more rigorously later. They are also helpful for feasibility testing at the early stages of shaping a new idea: looking at how something could work and getting a better picture of what's involved. Because they focus on one group only, pre-experimental designs share a common problem: without a control or comparison group it's hard to know if any change after the intervention is introduced was really caused by it, or by another factor affecting the group in question. When we observe a change in the world, like crime decreasing after the introduction of a new policing strategy for example, it can be hard to isolate the cause of the change. We might *hypothesise* that falling crime is the result of the new policy, but it could be the result of something else that was happening alongside the intervention – a *confounder*. This could be a change in poverty, the quality of police work, or just the time of year, that is influencing the outcomes we see. The control and comparison groups used in other types of experiments aim to mitigate the impact of *confounders*. For this reason, pre-experiments are usually unhelpful for telling us whether or not our idea is effective, but they can help us learn about the ingredients of a successful solution. If innovations developed with pre-experiments show promise, they can be evaluated with quasi-experimental or randomised designs.

Figure 3: Three basic experimental approaches[64]



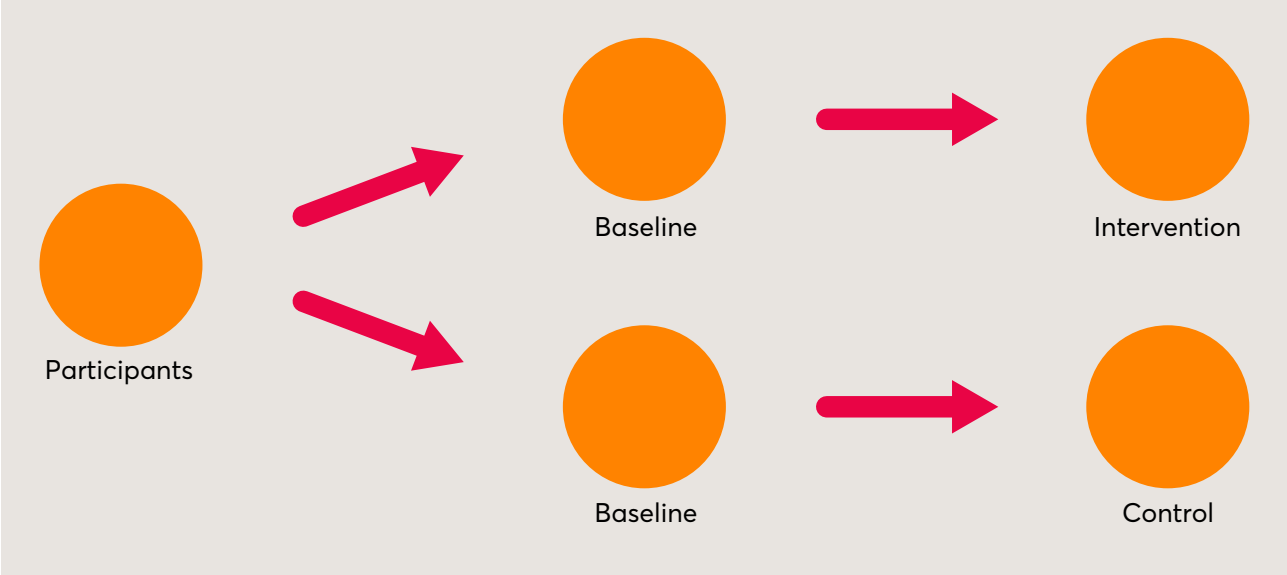
Source: Informed by Salkind (2010) and Salganik (2017)

As we'll see, there are many variations to these basic designs. In practice, experiments can have more than two groups – there could be three experimental groups and one control, for example. There can even be comparisons within one person – like the 'n-of-1 clinical trial' that compares different drugs and their side effects using an individual patient as the sole *unit* of observation.[65] There are also inventive ways of randomly allocating people to control and experimental groups if dividing the groups in a straightforward way isn't an option, or when you want to make sure everyone who is eligible can get the innovation; we cover some

of these below. One key design decision is whether an experiment compares intervention and comparison groups both before and after an innovation is introduced, rather than only after (see **Figure 4**). When possible, taking measures both before and after an intervention is introduced is recommended in experiments – this means we can take any differences between groups into account, instead of solely relying on randomisation to create a perfect comparison. We use the words 'innovation' and 'intervention' interchangeably throughout this guide – and you might occasionally see the word 'treatment', when we give an example of a new solution being tested in medicine.

The concepts in this section can be a lot to take on board. Check in with **Annex A**, our experimentation jargon buster, for a recap.

Figure 4: Taking before and after measures in a randomised trial



Section 3: An inventory of experimental approaches

This catalogue provides a whirlwind tour of experimentation. It's not intended to be a fully comprehensive list of experimental designs – our aim is to shine a light on new, important and interesting approaches being pioneered globally.

Table 1 below provides a summary to help government, charities, funders and foundations think about the options and choices available.

Table 1: Summary table: experimental USPs

	Design	USP	Pros	Cons
Randomised experiments **Best for increasing confidence about whether or not an innovation is making a difference** **Control group: Yes** **Randomisation: Yes** **Causal power: High**	**The basic RCT**	A straightforward design when you want to test an innovation against 'business-as-usual', and you have the opportunity to randomise who gets what.	Simple design, easy to explain to stakeholders Many guides available	Not always possible Requires you to have control of who gets what
	Multi-arm trial	A trial that tests multiple innovations in one experiment, to see which one is most effective, or to learn about the mechanisms driving the results.	Testing multiple interventions together can be more cost-effective than testing them separately Facilitates the study of mechanisms or comparisons of different theoretical predictions May allow for cost-effectiveness comparisons	Can be more complex or require more expertise Often requires a larger sample (and thus can be more expensive) than a single comparison
	Nimble trial	A test focused on short-term outcomes and operational questions, that aims to quickly get useful information into the hands of decision-makers.	Timely and useful information for decision-making Useful for formative evaluation (how to improve what you do) May be cheaper and faster than many other RCT designs	Not focused on ultimate impacts Often requires you to gather extra data such as surveys A newer approach in the evaluation community, so it may be harder to access support
	A/B Test	A test that compares two or more versions of a service or message, usually online, to see which works better for users and organisations.	Cost-effective High-quality tools freely available online Can be embedded into websites or platforms to improve performance and usability	Largely limited to online use Online tools may still require expert support to use well Can lead to 'tunnel vision', optimising your chosen outcome at the expense of others
	Cluster randomised trial	An experiment that allocates randomly chosen groups of people, rather than individuals, to an innovation. Helpful when individual randomisation isn't an option (such as changing the way a teacher teaches a class and measuring the effect on the pupils) – and when a policy idea aims to innovate at the level of the institution or area.	Allows you to learn about innovations at a local or institutional level Overcomes design challenges with individual randomisation May be more politically acceptable Helps limit spill-overs/*contamination*	Some design downsides, like higher costs and a bigger sample required If the *unit* of analysis (e.g. a pupil) is different to the *unit* of randomisation (e.g. a school) this can mean a slightly more complicated analysis
	Stepped-wedge and waiting list designs	A staggered experiment, in which the roll-out of a policy happens in stages and the control group does, in the end, receive the innovation. Helpful for overcoming political or PR difficulties in running a trial, and when we have good reason to believe that people will benefit from an innovation.	Often more politically acceptable Can have a dynamic model, where early learning informs later tests Everyone eligible gets the innovation being tested	Can be complex to organise Must be possible to control who gets what and when Only informative about short-run changes (as in the long run, both groups have received the innovation)

Table 1: Summary table: experimental USPs (continued)

	Design	USP	Pros	Cons
Randomised experiments **Best for increasing confidence about whether or not an innovation is making a difference** **Control group** Yes **Randomisation** Yes **Causal power** High	**Cross-over design**	A design in which each '*unit*' (usually a person) receives different innovations in sequence. As well as having a randomised control group, each individual acts as their own control: we can compare the relative impacts of different innovations, against business-as-usual, for a person over time.	Longitudinal design means that individuals act as their own control Comparing innovations in a sequence can be more efficient	Risk of 'carry-over effect' means it's possible that some of the effects of the earlier innovations might 'carry over' into later tests, and skew results
	Multi-site trial	Experiments that increase the evidence base by testing in more than one place. Results may be more generalisable, and trials are usually accompanied by qualitative research to understand any differences between how things work and are perceived across the different sites.	Learn more about what works in different places and contexts Learn more through a joined-up approach across locations Results can be more generalisable, as are less sensitive to unusual events that happen in any one site	Takes planning, coordination and lots of resources for research and evaluation Difficult to ensure the same quality and content of the interventions at the different sites
	Realist trial	An experiment that aims to create new theory about how and in what contexts an innovation makes an impact, by looking at the underlying mechanisms that generate outcomes.	Tells you more about how an innovation works and for whom Builds theory on what's likely to be effective in future	A new approach, limited expertise on running them May require more time and investment
	Hybrid trial	An experiment with two goals: to test an innovation and find out how it should be adopted in everyday practice.	Answers problems of routine practice Helps speed up implementation or adoption of new ideas Learns more by running two tests at once	Can have design limitations which could impact on getting reliable results
	Adaptive trial	An adaptive trial 'plans to be flexible'. Early results modify the course of the experiment, according to pre-planned criteria, making it possible to test multiple interventions or intervention variants more effectively.	A 'learn as you go' approach Can be used to test multiple interventions or variants of interventions May be more efficient, and hone in on effective treatments earlier Can be more cost-effective, as changes to how resources are allocated can be made mid-experiment	Complex to run, requires much statistical know-how and careful planning More stringent statistical standards may mean a larger sample size All adaptation must be planned (or results will be untrustworthy)
	Regression discontinuity design	An RDD can be used where an innovation is delivered to those who fall on one side of an arbitrary cut-off (such as being eligible for a certain benefit if your income is below a particular level). It compares those just above the cut-off with those just below it to estimate the effect of the innovation.	Robust method, good at linking cause and effect Many opportunities for use in policy and services	Doesn't tell you about impact for the whole population Only relevant where such a cut-off exists and isn't used for many interventions at once

Table 1: Summary table: experimental USPs (continued)

	Design	USP	Pros	Cons
Non-randomised and quasi-experimental designs **Use inventive designs and statistical techniques to create a comparison group to estimate the effects of an innovation. Particularly useful when randomisation isn't possible or desirable, or when a new policy is already in place.** Control group: Yes Randomisation: No Causal power: Medium	**Matching**	Useful when there's lots of data available but it's not possible to randomise – it creates a comparison group by 'matching' people who receive an innovation with similar people who don't. It's best if people had little or no control over whether they received the innovation.	Generates lots of useful information about people who might receive an innovation Can be used when policy rules and conditions make other approaches difficult	Requires some statistical expertise to run, as well as an evaluation team with good knowledge of the policy area being studied Relies on lots of relevant data being available and participants not self-selecting whether they get the treatment or not
	Difference-in-Difference	DiD comes in handy when a policy is introduced in one area or region, but not in another place that's similar or has comparable trends – and is on a similar trajectory pre-intervention. It compares the change after innovation in the treated area with the change after innovation in the comparison area (hence 'difference in difference').	Many opportunities to use this, such as devolved or regional policy differences Straightforward to understand and explain	Only works if certain assumptions are met, but they can be tested Requires data from before the experiment began, or a delayed start so it can be collected
	Synthetic Control	A data-driven approach useful to understand policy changes that have already taken place, by comparing many individuals, populations and places over time. It may be less susceptible to *bias*, since it creates its comparison group by blending many individuals or areas together.	Useful when the assumptions for a DiD aren't met Software freely available Proponents argue a transparent approach	A new approach, still being developed Requires excellent data and statistical know-how: not to be attempted without the necessary expertise
	Pre-post test	The simplest design, compares outcome measures before and after an innovation is introduced.	Simple to conduct and communicate Not usually very resource intensive, and requires less specialised expertise	Susceptible to different kinds of *bias*, like *selection bias*, which can limit the usefulness of results Not recommended when other approaches are possible
Pre-experiments **Explore what changes when an innovation is introduced, using only a single group. Useful for exploratory aims, formative evaluation (improving what you do) and trying out new ideas, rather than robust impact evaluation.** Control group: No Randomisation: No Causal power: Low	**Rapid cycle testing**	Iterative experiments useful for local problem-solving, that create a rapid feedback loop between testing, re-design and re-testing.	Useful for improving projects and programmes and solving local challenges Helps develop and flesh out new ideas, good formative evaluation (improving what you do) Continuous learning and improvement for teams Aims to generate useful results faster	Many different approaches, some designs better than others Not suitable for making confident estimates about impact
	Prototyping	An approach to trying out ideas in practice before an innovation is offered to recipients, to get feedback and input from stakeholders, improve the idea and increase its chances of success.	Good for developing new ideas Low cost and low resource Strengthens innovations through stakeholder input Creates products that may be easier for people to use and engage with	Many different approaches, some over-promise on what they can deliver Doesn't allow for proper testing, more about refining ideas Some approaches don't help us learn from available research and evidence

Randomised experiments

USP: Best for increasing confidence about whether or not an innovation is making a difference

Randomised controlled trials, or RCTs, randomly allocate participants to control and intervention groups. Randomisation creates groups that are comparable before the innovation, which means that any group-level differences we observe afterwards can reliably be attributed to the innovation. Unlike some other kinds of experiments, they allow us to make stronger claims about cause-and-effect – they are especially good for drawing the kind of conclusions that researchers call *causal inferences*. That is, to say that we believe an innovation or solution is responsible for a change in the world. As the Government's Magenta Guide explains, an experimental impact evaluation:

> "[When] conducted properly, will establish whether an intervention caused an outcome... this is the strongest form of design for an impact evaluation, as the random allocation minimises the likelihood of any systematic differences – either known or unknown – between the groups. It therefore allows for an attribution of cause and effect."[66]

While RCTs are mainstream in medicine, there has been a recent boom in the number taking place in social policy areas. In education, there were only around ten RCTs published each year in the early 2000s, but this had grown to more than a hundred a year by 2012.[67] The UK's Education Endowment Foundation has now conducted more than 180 trials. Between 2003 and 2012, the number of trials in social work rose from ten a year to more than 50.[68] In policing, around 400 trials have now been run – with 40 new studies being added a year.[69]

We must be clear however that RCTs are not a panacea: it's perfectly possible to have a badly designed RCT, or one in which randomisation doesn't work to reduce *bias*. But, the randomised trial is the best tool we have for exploring cause-and-effect, and evaluating the impact of our attempts to change the world for the better. The following pages cover a range of different types of randomised experiments. In practice, there are many overlaps. A 'nimble RCT', for instance, can also be an 'A/B test'. We have classified different kinds of trial according to their different uses and characteristics.

3.1 The basic RCT

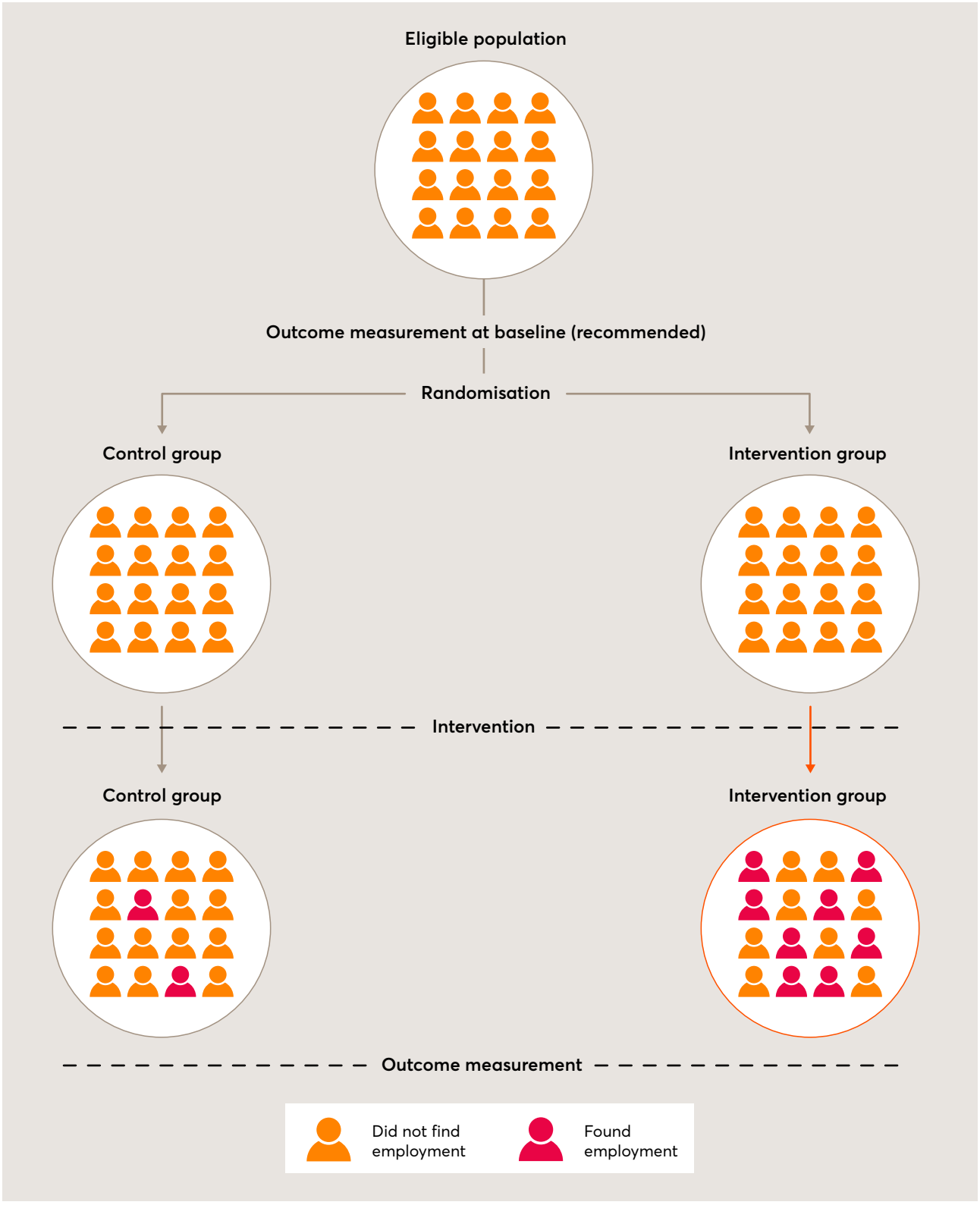
USP: A straightforward design when you want to test one or more innovations against 'business-as-usual' – and have the opportunity to randomise who gets what.

The introduction of a randomised control group reduces a whole host of biases that normally complicate the evaluation process. Let's take the example of a new 'back to work' scheme. How will you know whether those receiving the extra support might not have found a job anyway? Perhaps the economy got better, or there could be another factor at play that we aren't even aware of. **Figure 5** shows how the method works by randomly allocating some people to receive the scheme, and others to a control group. Measures taken before and after the innovation help us to compare how differences in outcomes between the two groups change during the experiment. By comparing outcomes for the control group and the experimental group, we can find out if the 'back to work' intervention helped people find jobs.

Separating out the effects of an innovation from other factors is the central aim of an RCT – and what gives them particularly strong *internal validity*, the technical term for the ability of an experiment to make cause-and-effect claims about the particular group in the study.

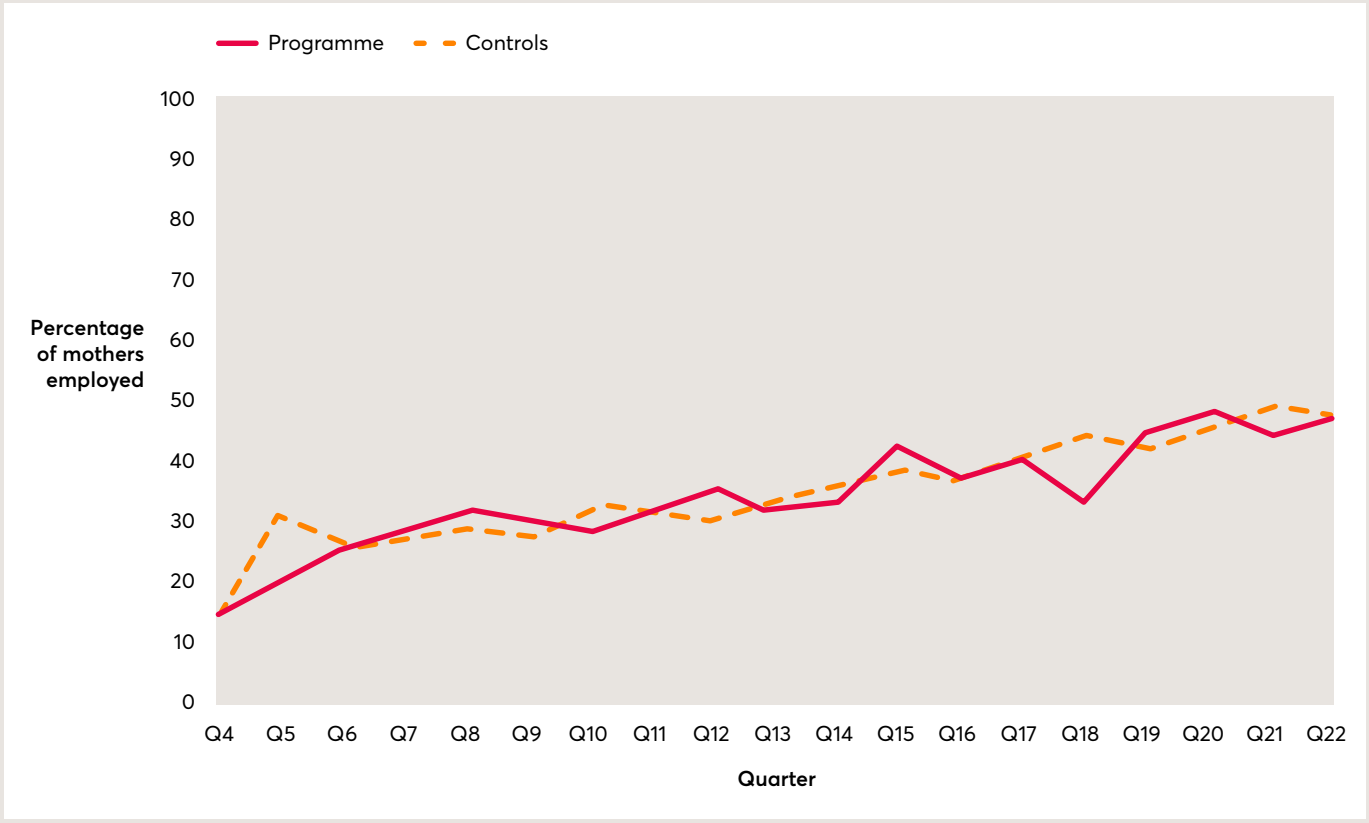
This strength of randomised trials has important policy implications. To take an example, during the 1990s, the Comprehensive Child Development Program was set up by the US federal government to help low-income families with infants or young children. The programme used specially trained staff to conduct home visits and provide case management services to participating families, with the goal of improving family economic wellbeing and children's development. In **Figure 6** below, the solid bold line shows economic outcomes of the 2,200 mothers who participated in the programme – specifically, the growth in their employment rate over five years. This might be regarded as a success: there was an impressive rise of employment from 16 per cent to around 40 per cent. However, the dotted line shows employment outcomes for a randomly assigned control group of 2,200 mothers who didn't participate in the programme. This control group experienced an almost identical rise in employment. Change had happened, but the Child Development Program wasn't responsible for it.[70] Examples like this highlight the benefit of rigorous evaluation – randomised trials ask and answer some of the toughest questions about policy decision-making.

Figure 5: Using a randomised design to test a 'back to work' scheme



Source: Adapted from A guide to randomised controlled trials, Innovation Growth Lab, 2016

Figure 6: Evaluation of the US Comprehensive Child Development Programme



Source: Jon Baron

Randomised experiments can work to different timescales. One of their benefits, although also a source of criticism, is that they allow us to investigate the effects of policy change over several years, to find out whether any positive impacts take a while to emerge or stand the test of time. An experiment in industrial policy conducted by Nesta in 2009 highlighted this (see **Box 3**). Initially positive results faded over the course of the year, making the scheme for boosting creativity among small businesses a poor investment over the longer term.

Box 3: Creative Credits: an industrial policy experiment

Nesta used an RCT to see if a novel business support scheme connecting small businesses and creative providers to boost innovation was effective. The pilot study, which began in Manchester in 2009, was structured so that vouchers, or 'Creative Credits', would be randomly allocated to small and medium-sized businesses applying to invest in creative projects such as developing websites, video production and creative marketing campaigns, to see if they had a real effect on innovation. The research found that the firms who were awarded Creative Credits enjoyed a short-term boost in their innovation and sales growth in the six months following completion of their creative projects. However, the positive effects were not sustained, and after 12 months there was no longer a statistically significant difference between the groups that received the credits and those that did not. Nesta published a report on the study, which argued that these results would have remained hidden using the normal evaluation methods used by government. The experiment spurned much further research on how to boost creativity using innovation vouchers internationally.[71]

One criticism of RCTs is that they can be expensive and time consuming. RCTs do require resources, but the traditional image of them comes from R&D in medicine – where full-blown drug trials can take ten to fifteen years to get from lab to hospital, with three painstaking stages of clinical trialling.[72] Randomised experiments can however be swifter, cheaper, and better adapted to the real world of a hospital, school, or neighbourhood. As we'll explore in **Box 8** on online experiments, online delivery can help minimise costs. There is nothing inherently expensive in a trial compared to other types of evaluation. The resource-intensity often comes from the innovation (such as a big new welfare programme) or gathering data and insight for the evaluation itself (such as surveying thousands of people). But little of this burden is unique to a randomised trial.

Most randomised trials are now also conducted alongside a 'process evaluation'. This is qualitative research that explores how the innovation being studied aims to bring about change, and what the practical work of doing the innovation means for the people (whether that's teachers or social workers) who will deliver it. Process evaluations are now strongly recommended by the Medical Research Council as an indispensable part of any evaluation of a complex intervention.[73] They help us understand why something has or hasn't worked, and build reliable 'theories of change'[vi] about how an innovation aims to make a difference through its activities and outputs. All trials run by the Education Endowment Foundation (EEF) for example include implementation and process evaluations, which map out a theory of change for the innovation being tested.

Box 4: Efficacy vs Effectiveness

Trials can be down-to-earth. 'Efficacy' and 'effectiveness' are two different approaches to running an RCT, and prioritise learning different things about an innovation, summarised in **Figure 7**. Efficacy research is focused on whether or not an innovation works under ideal conditions (a very controlled situation, like a lab or a classroom hosting a research team). On the other hand, effectiveness research focuses on whether an innovation works in real life, and at scale. Each of these has its costs and benefits: an efficacy test might tell us more about cause-and-effect relationships, which would help us to identify other contexts where similar interventions might be effective. An effectiveness test would allow us to try it out with more people and observe any side effects or issues with implementing the innovation in a real-life working environment, like the everyday classroom, town or hospital.

Figure 7: Efficacy vs effectiveness

Effectiveness trials involve much larger samples than efficacy trials, which focus on conducting a more strictly *controlled experiment*. This control makes efficacy experiments strong in terms of *internal validity* – that is, producing a fair and unbiased representation of the treatment effect within the experimental sample. Effectiveness trials do better on *external validity*: that is, the results should generalise well, by involving diverse people from the general population, and looking at an innovation in a more natural context.

In reality, efficacy and effectiveness are perhaps best seen as a continuum, one that provides a useful way of thinking about some of the trade-offs involved in experimental design. In public policy, most trials have to find compromises to address these tensions. Researchers will consider how to select a sample of people to take part, how large and diverse that sample should be, and how they go about ensuring that the innovation being trialled is implemented as its designers intended (what's called the *fidelity* of the trial). In social policy, most innovations can't be developed in a lab – new policy and practice solutions must always work out in the world to be considered effective.

It is important to acknowledge that RCTs can still suffer from their own substantial biases.[74] Some of these can be controlled with careful planning, but others can't be fully mitigated and must be considered when interpreting their results.[75]

In some instances, trials may be unworkable. It is impossible to randomly allocate some variables, such as age or sex. Additionally, some variables would be unethical to randomly allocate. A study of the long-term effects of nutrition or poverty would not randomly

allocate people into nourished vs starving, or poor vs rich. **Annex B** gives a summary of some common criticisms of RCTs and responses to them – adapted from the National Foundation for Educational Research.[76]

An important critique of randomised experiments comes from Professors Angus Deaton and Nancy Cartwright – who argued in 2016 that there are several 'misunderstandings' of randomised controlled trials, which result in researchers overclaiming about what they can achieve.[77] In particular, they point out that randomised trials can have limited *external validity* – that is, they are not always good at producing generalisable knowledge about 'what works' in different contexts, and for different people (although this may be equally true of other kinds of research). Organisations like JPAL, an international network that works to reduce poverty through experimentation, have developed practical tools to help address these challenges.[78]

There are also public concerns about the ethics of social experimentation. For people who are unfamiliar with how trials work, or unaware of what design options exist, running experiments on essential services can seem objectionable. But, some of these issues are based on misunderstandings – as well as a failure to consider what the cost of not testing policy could be. **Box 5** covers some of the ethical arguments about an experimental approach to problem-solving.

Box 5: The ethics of randomised experiments

The ethics of randomised experiments has been hotly debated. One public concern is that it might be unethical to experiment on humans – and to deny people an intervention that could help them. A common question here is 'why give a treatment that can help one group but deny it to the control group?'[79] We have argued that we can borrow from medicine to help us navigate this dilemma safely. The principle of medical equipoise states that *"trials can only be justified if there is genuine uncertainty in the expert medical community about the preferred treatment. A physician must have an equal state of uncertainty – or 'equipoise' – between the available options."*[80] So, we shouldn't deny people the chance of benefiting from innovations that we have good reason to believe will help them.

There are clear cases where it would be madness to experiment, like with interventions that could cause people harm, for example testing how people respond to smoking cigarettes.[81] In these circumstances, running an experiment would be unacceptable. But there are also equally clear cases (and examples from history) in which not experimenting can cause significant harm. A classic medical example was giving steroids to people with head trauma; for years an RCT was not approved because it was deemed unethical. But when one was finally carried out it was found that the steroids we'd been giving to patients for years were actually increasing their mortality – and not running a trial had resulted in preventable deaths.[82]

In social policy too we must consider the cost of not doing a trial and rolling out a policy that is untested, and potentially doing harm or wasting taxpayers' money. Effective innovations can also have side effects and a trial can be a cost-effective way of identifying these, so that future implementation plans can build in ways to mitigate them.

There may be cases when it's less clear cut whether or not to run a trial.[83] We should be guided by ethical principles in both how we invest in and evaluate social policies. Finnish government has led the way in establishing an ethical code of conduct for their national programme of policy experiments.[84]

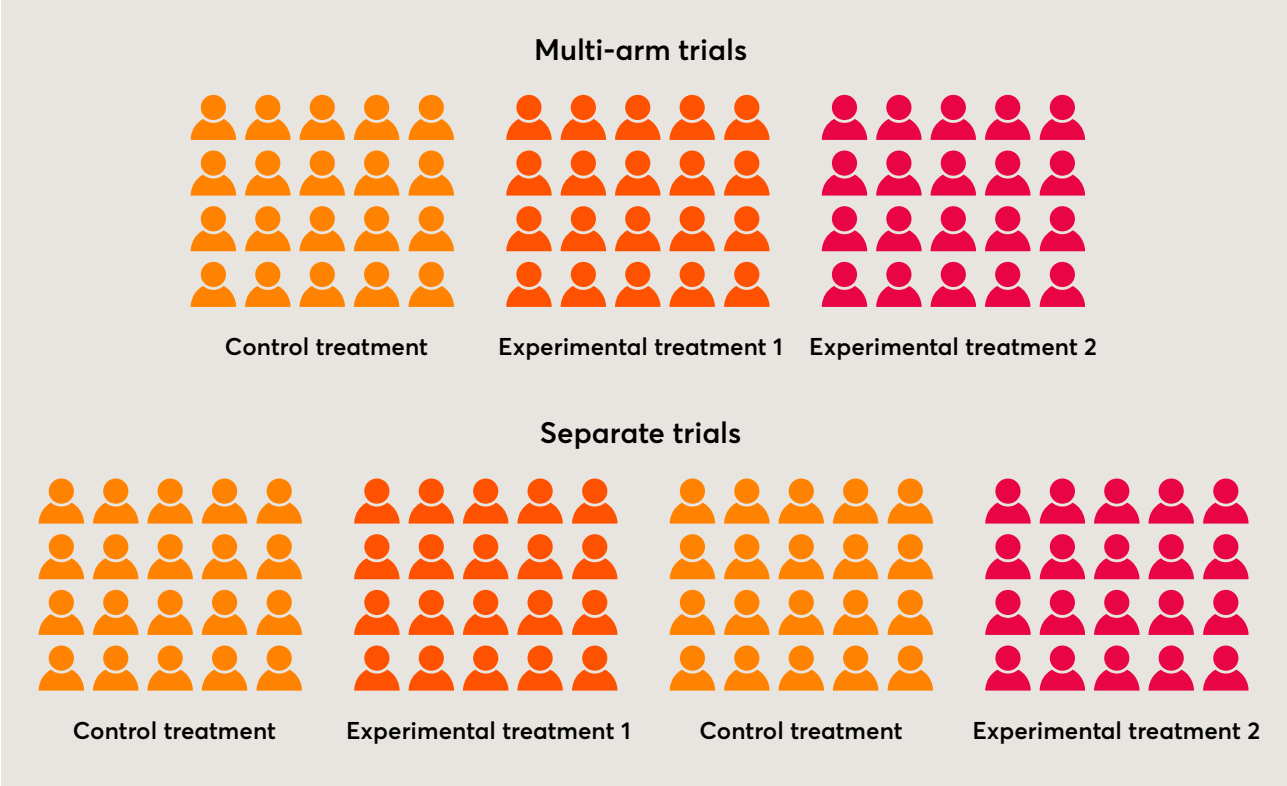
There is a large and valuable body of academic literature on RCTs, including how we can learn most from them (see **Annex B**). The different designs we describe below aim, in different ways, to address some of the limitations of the basic design.

3.2 Multi-arm trial

USP: A trial that tests multiple innovations in one experiment, to see which one is most effective, or to learn about the mechanisms driving results.

Trials can also test a variety of things at the same time: different versions of an innovation, or variants of a service. Multi-armed trials aim to speed up how we get evidence about these different options. Instead of doing lots of individual studies with one intervention and one control group, different versions of an innovation can be compared against a common control group in a single trial.[85] As **Figure 8** shows, running separate RCTs will require a greater number of individuals in a control group, compared to a multi-arm trial with a shared control group.

Figure 8: A multi-arm trial and the benefit of a shared control group



Source: Adapted from *Trials* Watson et al, 2017

In medicine, the UK Medical Research Council (MRC) Clinical Trials Unit estimates that doing multi-arm studies could cut the costs of experimental research by half. They have advocated for multi-arm, multi-stage (MAMS) trials that take this one step further.[86] These not only test several treatments at once but are also more responsive to early results – you

can stop recruiting patients to a particular arm part-way through, if a treatment is not looking promising (for more on these, see Section **3.11**). They even allow experimenters to add new treatments to the trial when they are ready for testing. According to the MRC Trials Unit:

> *"This approach is even more efficient, meaning you don't have to have gaps between stopping a trial and starting a new one, and may not need to set up a separate new trial to test a new treatment. This saves many years… One example of this is the STAMPEDE trial, looking at how to treat prostate cancer. It has stopped and added arms as it's gone along. It now has more than 5,500 participants, making it the largest prostate cancer treatment trial ever. It will assess eight new treatments in 15 years – something that would have taken at least 40 years in separate sequential trials."*[87]

A further benefit of multi-arm designs is that because they compare different innovations, they can help us understand the strengths and weaknesses of different ways of trying to create change and improve outcomes, and build on the theory that underpins them. Multi-arm trials have even greater potential for improving the efficiency of experiments when they are combined with machine learning. Multi-armed bandit experiments (**Box 6**) use artificial intelligence to learn from early results and tweak the number of people allocated to different intervention 'arms'.

Box 6: Multi-armed bandits

Multi-armed bandit (MAB) experiments are so-called because they are inspired by Las Vegas gambling machines, which have two arms that a player pulls to try and win cash. When playing a whole casino floor of these slot machines, some may have higher hit rates than others. But the player would only find that out by playing all of them. Once you know which machines are the winners, you can focus your attention on them and would have a better rate of success.

Thanks to machine learning and mathematical theory, this method can be applied to experiments that are run in a sequence in different places and different times: more people can be added to intervention arms that look successful, or be removed from those that don't. It's a learn-as-you-go approach. For instance, if you are running a monthly training programme across an entire company, you could add more people to the most effective types of training as the experiment progresses. Complex mathematical approaches will be needed to help you, but the benefit could be that you don't waste time randomly allocating lots of people to an approach that doesn't work. For something like a new breast cancer drug, this is an important point: you can get effective therapies to patients quickly and avoid assigning patients to ineffective treatments.

MAB trials have attracted more interest in the digital domain, as a more sophisticated way of doing A/B testing (see Section **3.4** for more on A/B tests),[88] and are yet to take off in other fields. For instance, despite much discussion of this technique in medicine, MAB experiments are rare in clinical practice.[89] MABs are likely to only be useful in very specific circumstances. However, there is more potential for such trials on social issues, such as public health campaigns or e-learning.[90] For now though, MABs are a novel and advanced technique. If you are without specialised support, we would advise you don't try this at home!

3.3 Nimble RCT

USP: A test focused on short-term outcomes and operational questions, that aims to quickly get useful information into the hands of decision-makers.

A pragmatic approach to experimentation highlights the need for useful, applicable testing. Decision-makers require research that is fast, flexible, and cost-efficient – and that responds to the problems and questions most pertinent to organisations. In international development, Dean Karlan, head of the NGO Innovation for Poverty Action, coined the phrase 'Nimble RCTs'. These are randomised trials that test tweaks and changes to how innovations are delivered. They focus on short-term outcomes and operational questions, so they are not primarily geared to measure impact. One nimble experiment run by Innovations for Policy Action was a collaboration with Telenor Pakistan, a mobile money provider, that aimed to increase the use of its platform Easypaisa by the country's poorer citizens. They tested marketing campaigns, offering different kinds of incentives, and targeted referral programmes, looking at ways to get more sustained engagement.[91] These small tweaks to services can lead to big improvements in efficiency. Although they aim to get results fast, they are applicable across many policy areas and form a central part of the portfolio of experiments that Nesta's new **EdTech Innovation Testbed** is running, which will help schools and colleges to trial promising technology, as part of a partnership with the Department for Education to support more effective use of tech in schools.[92]

The World Bank is also exploring this flexible approach to trials; in 2018, its Strategic Impact Evaluation Fund (SIEF) put out a call for 'nimble plumbers' to work on rapid, low-cost evaluations that generate experimental evidence on implementation and delivery issues.[93] They funded nimble evaluations on how best to improve the take-up of health insurance in Azerbaijan, expand the use of contraceptives in Burundi, and support teachers to deliver tailored education to children affected by war and displacement in Lebanon.[94]

Box 7: Three nimble trials run by the World Bank

1. **Burundi**: Expanding contraceptives use when women are shunned for using them

   The Government of Burundi wants to reduce the fertility rate to three children per woman, from the current rate of 5.5 children per woman. Because there is a strong opposition to birth control from faith leaders, and because women who use birth control risk social isolation as a result, the government wants to introduce self-injectable contraceptives. But getting community health clinics ready to prescribe self-injectables and ensuring privacy for women will be challenging. Using data from Burundi's health monitoring and information system, this evaluation will test different approaches for increasing adoption of these contraceptives. (Researchers: Arndt Reichert)

2. **Ecuador**: Increasing entrepreneurship and STEM careers through an online course in high school

   Entrepreneurship in high-growth sectors and employment in sectors requiring training in science, technology, engineering, and maths (STEM) offer an opportunity for high-paying careers, but most young people in low-income countries do not select these educational and career paths. In high school, students may not be aware of these career paths, they may lack the appropriate skills, and they may lack role models. This evaluation will test the impact of offering a 12-week online course in high school on student beliefs about and interest in high-growth entrepreneurship and STEM careers. Courses include personal initiative and negotiation skills, basic scientific methods, and role model interviews with entrepreneurs and scientists. Using data from an online monitoring system designed in tandem with the course, the evaluation will also test strategies for increasing course adoption among teachers. (Researchers: David McKenzie, Igor Asanov, Diego d' Andria, Mona Mensmann, Bruno Crepon, Guido Buenstorf, Tom Astebro)

3. **India**: How can we reduce medical billing errors?

   Evaluation: Health systems increasingly rely on administrative payment systems to reimburse service providers for claims they submit to insurers or governments. In these systems, billing errors can threaten policy objectives and the viability of public programmes, in addition to causing patients to lose money. Errors may be mistakes, or they may indicate waste, abuse, or fraud. The issue is of great importance in India, which is in the process of unrolling a universal insurance plan for hospital care to 500 million people. This evaluation, which will rely on administrative data collected through the insurance programme and cross-referenced with medical records and patient information, will test the impacts of providing hospital management with private 'report cards' on their billing errors and guidance on reducing errors, informing patients about claims filed in their name, and a combination of the two interventions. (Researchers: Sebastian Bauhoff)

Source: Adapted from World Bank, SIEF brief, Nimble Evaluations available at worldbank.org.

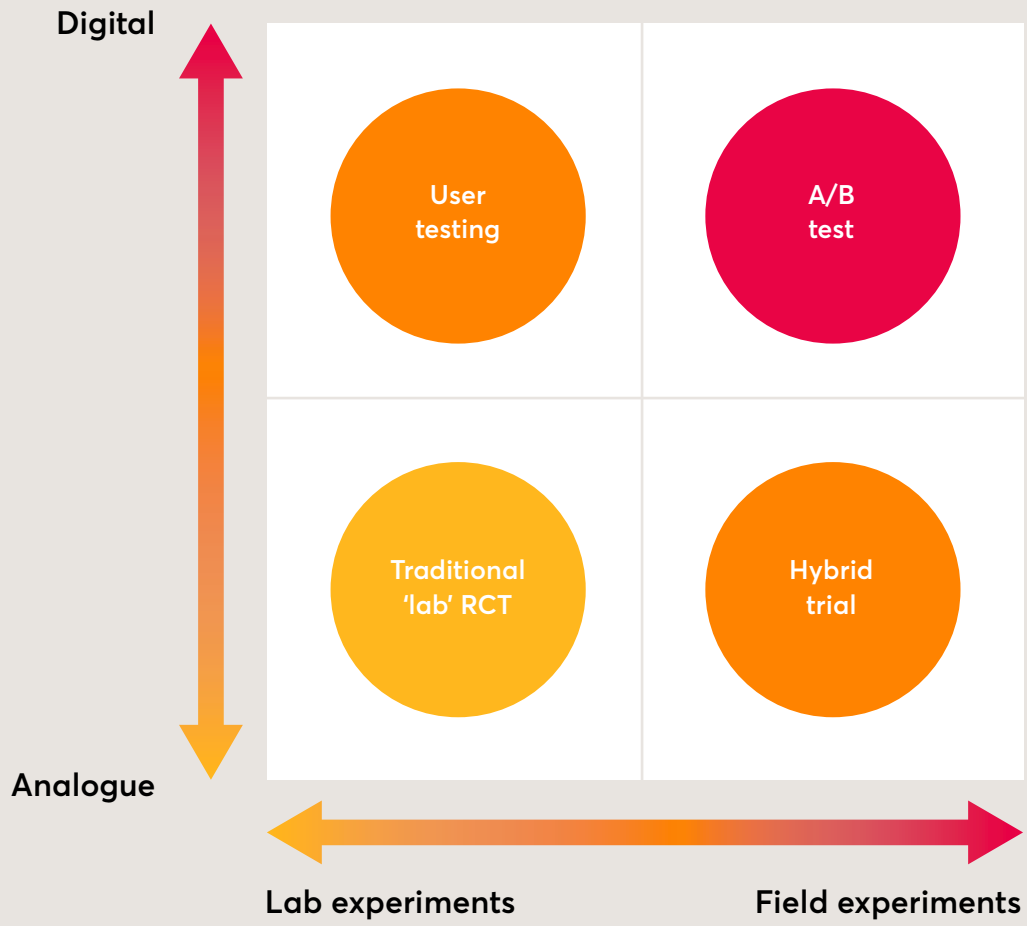
Mary Kay Gugerty and Dean Karlan's recent book *The Goldilocks Challenge* points out that nimble trials are an important part of getting 'right fit' evidence for organisations.[95] They can be especially useful for answering questions about the early stages of a theory of change, a visual map of how an innovation works, to ensure services are efficient and well-designed. These trials are focused on do-ability and timeliness, generating evidence that's immediately useful for the day-to-day work of improving how innovations are managed.

In some cases, where randomisation isn't possible, 'nimble' testing can also be done using a quasi-experimental comparison group (see Sections **3.12-3.15**). While nimble trials are likely to be substantially cheaper than traditional ones, they still require a large enough sample size to get reliable results and may involve collecting additional data, for example through a survey. Nimble RCTs are sometimes called 'rapid-fire tests' and, when done online, can be called A/B tests.

Box 8: New possibilities for digital and online experiments

Technological change is transforming the landscape of experimenting. Traditionally, experiments have fallen into two categories: 'lab experiments', conducted in specially designed environments; and 'field experiments', run out in the world. However, according to the Princeton University computational social scientist Matthew Salganik, we can now add another dimension to where and how experiments happen: 'digital experiments' versus 'analogue experiments', that take place 'offline'.[96]

Figure 9: Where do experiments happen?



Source: Adapted from *Bit by Bit*, Salganik, 2017

One advantage of digital experiments is that they can dramatically reduce the cost of running trials. Online experiments can be cheap and comparatively easy to run, particularly if they are making a tweak to an existing online service. We can experiment with small changes to government websites at low cost and reach large numbers of people. In fact, this process can now be automated entirely. The Government Digital Service and the Behavioural Insights Team have experimented to improve everything from encouraging organ donations (see **Box 9**), to filling in HMRC tax forms.[97]

Web-based experiments can be unique. Compared to 'analogue' experiments, they may be statistically more powerful if they draw on access to larger samples online. Arguably, online samples are more representative of the national population than the typical western, educated, industrialised, rich and democratic (or WEIRD) college student participant used in a traditional academic trial.[98]

Online research has grown exponentially in the last 20 years, and with the advent of easy-to-access resources such as MTurk (Amazon's Mechanical Turk), a crowdsourcing marketplace for recruiting online participants, it shows no signs of slowing down.[99] Some of the experiments can be fun for participants, as well as providing rich data for researchers. LabInTheWild provides users with feedback, letting them compare themselves to people around the world while also contributing to cross-cultural psychology and social science.[100]

There are a plethora of different platforms for running online experiments, including Qualtrics, Gorilla, Wextor, and Testable. The University of Deusto in Spain curates a list of more than 700 psychology web experiments used by researchers from many universities.[101] Some of them can test very complex tasks, such as reaction time data with millisecond accuracy for cognitive tasks, and game-like tests such as Tetris and Tower of Hanoi. They can track your use of mouse, keyboard, voice recording and basic eye-tracking. Researchers from a range of fields, like psychology, clinical research, population health, social sciences, economics and more, are turning to the web to supercharge their science.

Experiments may also be more realistic if done via the web – closer to real human interactions. Experimental sociologist Damon Centola of Pennsylvania University describes in his book *How Behaviour Spreads* a test to create digital health communities to study 'complex contagions' such as social movements, political campaigns, or – in this case – how people adopted healthy behaviours.[102]

Other experiments can aim to mimic everyday life. Some take place on platforms where users are already online, so it is not a fake environment for them but a digital 'place' where they were already spending time. For example, a massive experiment on Facebook (with a sample of 690,000 people) showed how people's emotional states can be transferred to others via 'emotional contagion', leading people to experience the same emotions as other people online, without knowing.[103]

Internet research does also however face a unique set of problems. It has higher dropout rates, can exclude sections of the population who engage less with technology, and there is also the possibility of repeated participation.[104] There are also serious concerns that data may be poor quality, because online participants are less attentive or less motivated[105] – a problem Amazon's M-Turk aims to mitigate by paying people to join trials. There are of course limits to what can be tested online, and some of the advantages listed here can cut both ways. Online environments may not successfully mimic real-life ones; and asking people about their preferences online isn't the same as seeing what they do in the real world. But, as the core services of government and organisations are increasingly digitised, opportunities for cheap experiments with big impacts are growing.

3.4 A/B test

USP: A test that compares two or more versions of a service or message, usually online, to see which works better for users and organisations.

Some people claim that the rise of 'big data' means we may no longer need randomised experiments. Instead, they argue, large volumes of data will mean we can just look for patterns. The advent of big data will transform how we learn about social and public problems – and we point to some uses of new data sources in Sections **3.12-3.15** on quasi-experiments – but it doesn't equal the demise of the randomised trial. As the politician and economist Andrew Leigh points out in his book *Randomistas*, Google has arguably more data than any other organisation in the world – around 15 billion gigabytes at the time of writing and increasing at a rapid rate – but still conducts randomised experiments.[106] They, and other businesses such as eBay, Chrysler, United Airlines and Uber, run experiments called 'A/B' tests, which have become a central part of the day-to-day operation of internet-based companies. Quora, a 'question and answer' website, conducts around 30 experiments at any given time.[107] A/B tests work by allocating users to either an experimental or a control group, and analysing what messaging, website layout, or communication strategy works best. Tests that involve more than two options, other than A and B, are called 'multivariate' or 'split' tests.

Figure 10: A/B testing online

Source: Adapted from Rachuri, 2017

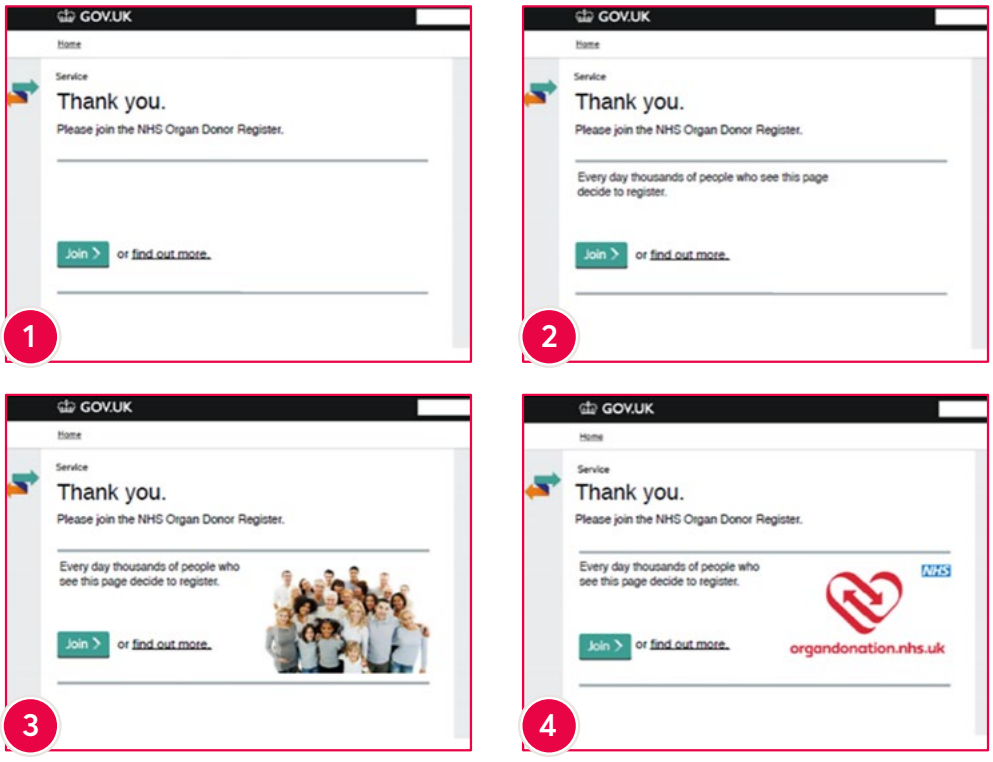
Any time you surf the internet, many experiments are being performed on you. The content and design of websites are varied randomly to figure out which versions make you most likely to engage in the intended way. Without being told, a group of users are diverted to a different version of a given webpage. Their behaviour can be compared against the mass of users on the standard site. So, if the new version gains more clicks, longer visits or more purchases, it will replace the original. If not, it can be quietly phased out. A/B testing is also used in many other areas, such as online donations for charities, encouraging volunteering, or political campaigning. Obama's successful 2012 presidential campaign ran more than 500 A/B tests that boosted sign-ups to his campaign by 161 per cent.[108]

A/B tests are not always one-off experiments, but can adapt, grow, and exploit the power of machine learning. As algorithms learn more about you and people like you, they adapt their behaviour in line with what they discover and feed this learning into future tests. Such algorithms, which adapt their behaviour over time, go by the colourful name of 'multi-armed bandits', and are covered in **Box 6** earlier in this report.

Box 9: The organ donation trial

The UK's Government Digital Service has used 'split' tests as part of its efforts to digitise the administration. In partnership with the Behavioural Insights Team (BIT), it ran one of the largest randomised trials ever conducted in the UK, with more than one million people. This experiment trialled eight different versions of the NHS Organ Donor Register webpage to see which was most effective in encouraging members of the public to join the register. The trial found that even small changes made a significant difference; if it were used nationwide for a year, the best-performing option would lead to 96,000 extra registrations, compared with the control condition. [109]

Figure 11: Four of the nine website variants used in the organ donation trial



Source: Behavioural Insights Team

Behavioural insights *units* – or so-called 'nudge *units*' – are taking advantage of opportunities to experiment in online systems, improving and evaluating solutions using evidence from behavioural science. The Behavioural Insights Team in the UK, NIT in the Netherlands, BETA in Australia, and TEN (the European Nudging Network) are just some of the organisations applying insights from academic research in behavioural economics and psychology to public policy and services. The Behavioural Insights Team was the world's first government *unit* to use digital experiments to tackle policy challenges. The team has implemented low-cost, high-impact changes in fields as diverse as taxation, healthcare, employment and sustainability. And, they have pioneered the use of simple online randomised trials in policymaking. One such test, run in partnership with the UK's Government Digital Service, was one of the largest ever trials conducted in the UK, involving more than a million members of the public (see **Box 9**).[110]

The organ donation trial shows one of the major benefits of digital experiments: access to large samples of experimental participants, often at no additional cost. Digital experiments also promise other advantages, like pre-existing and 'always on' measurement systems that expand the number of things experiments can test for.[111] For example, in the Facebook experiment we mentioned earlier (see **Box 8**), researchers would have access to lots of pre-existing information about individuals' backgrounds and demographic characteristics, likes and preferences, and social history online. Because of the lower cost of digital experimenting, it's often possible to run trials for longer, as well as in multiple (geographical) places at once.

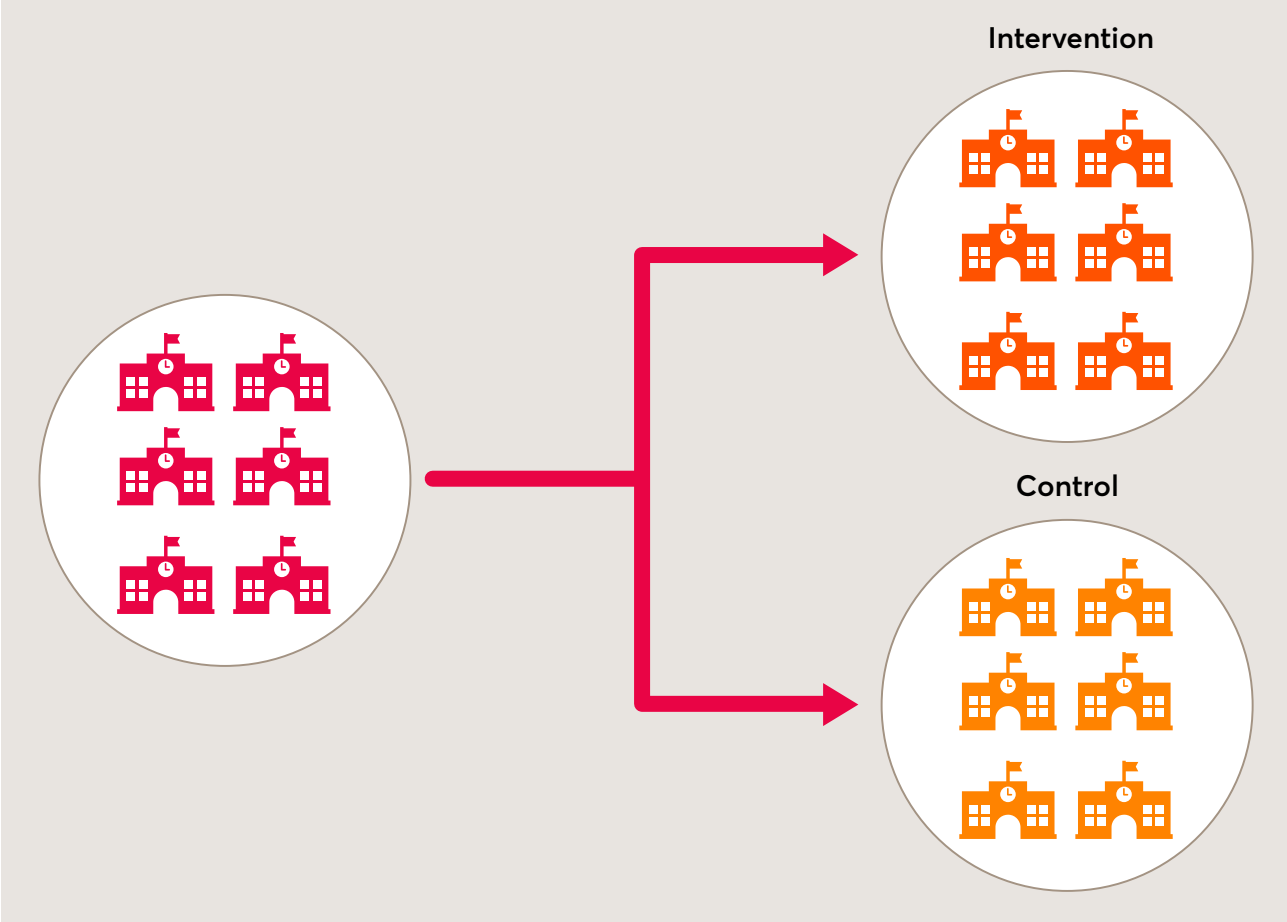
A/B tests aren't always well designed. There are lots of options for doing 'DIY' online testing but some of those websites don't generate results based on reliable sample sizes that would allow for proper statistical analysis. Options like Test+Build and Predictiv, created by scientists at BIT, offer more robust tools for organisations.[112]

3.5 Cluster randomised trial

USP: An experiment that allocates randomly chosen groups of people, rather than individuals, to an innovation. Helpful when individual randomisation isn't an option (such as changing the way a teacher teaches a class and measuring the effect on the pupils) – and when a policy idea aims to innovate at the level of the institution or area.

Randomisation can happen at an individual level (where people are randomly assigned into control or experiment groups) or at a group level (when pre-existing groups such as schools, villages or hospitals are randomly assigned to control or experiment groups). This second type, where randomisation happens at a group level – like a hospital, neighbourhood, or geographical area – is called a cluster randomised controlled trial (also known as a group-randomised trial or place-randomised trial), shown in **Figure 12** below.[113] Cluster trials are helpful when it's hard to randomise at the individual level, but can still be used to analyse individual-level outcomes. It would be difficult, for example, to compare two methods of teaching a topic, or two reading schemes, to pupils sitting next to each other in the same classroom. *Contamination* might arise, as one half would hear what the other is being told, and vice versa. To get around this source of *bias*, it may make more sense to randomise at a higher level – of whole classes, schools or catchment areas.

Figure 12: Random assignement of groups in to intervention and control



Source: Adapted from *Understanding research methods*, Parliamentary Office for Science and Technology (POST), forthcoming 2019

Randomisation can be politically tough to sell to voters: some constituents may be outraged that they are not receiving a new service offered as part of an RCT. The reality of democratic politics is that randomisation can create controversy. In Project STAR, a large US education experiment designed to test the effects of class size, about 10 per cent of students were moved to classes of different sizes than the ones to which they were randomly assigned at first, in part because of parental complaints and organised lobbying.[114] A school-level cluster design might have mitigated against this problem; all the children within a school would have been subject to the same rules during the experiment, so there would be less for concerned parents or lobby groups to object to. There are also ethical arguments that support cluster randomisation. Imagine randomising free school meals within a school for example. This would quickly become both practically and ethically untenable. A cluster design can help to reduce some major sources of *bias*, such as *contamination* (in our school meals example, imagine what happens if students share their food with friends). Another is experimental effects – if we gave free meals to some students within a school and not others, teachers might intervene, deciding to help out those students in the control group. Frustrated citizens may take matters into their own hands.[115] Cluster trials can be a good way to overcome these challenges, as well as being the most sensible level at which to test many kinds of innovations.

One example of a cluster design is an experiment on how the Mexican Government extended healthcare to over half the population. The Seguro Popular de Salud (Universal Health Insurance) was evaluated through a phased and random implementation, and is regarded as the largest ever randomised health policy experiment, involving more than 50 million Mexicans who previously had no health insurance.[116] It was a major part of the 2006 Mexican federal elections and much political capital had been invested. The federal government spent the equivalent of US $795 million on the new policy in 2005, entirely new money spent on the health sector. A research team from Harvard was tasked with an evaluation at the request of the Mexican Ministry of Health. Seventy-four clusters were matched in pairs so that one received the intervention and the other acted as control. In this particular case, a commitment was made to make the innovation available to control clusters on completion of the study.

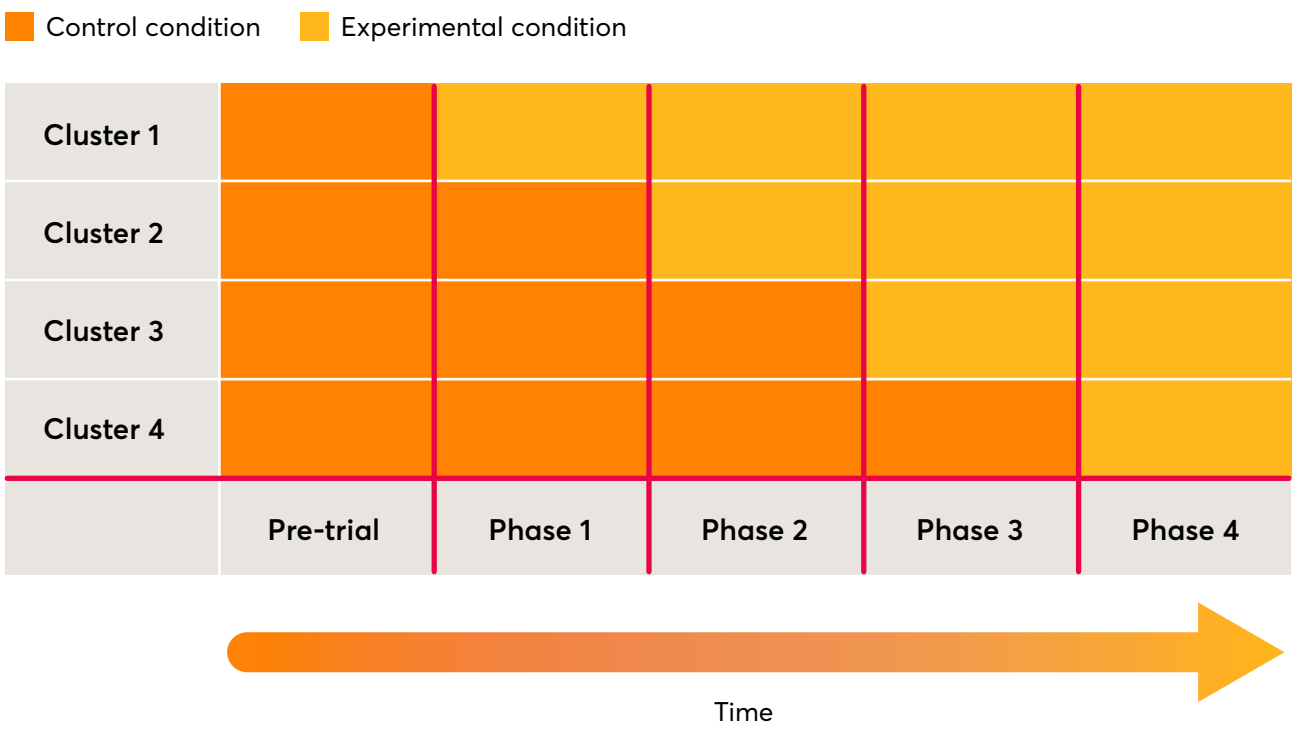
The level of random assignment chosen by the evaluators was named a 'health cluster'. It included a health clinic and the catchment area around it. The cluster experiment was politically feasible because the policy was always going to be rolled out to different parts of Mexico gradually, not everywhere at the same time. The Harvard team was thus able to randomly choose which area got the new policy and evaluate its impact. The Mexican healthcare cluster trial was a great example of how to politically inoculate an experiment. The use of an independent evaluation team also helped to embed the policy beyond the political cycle of the incumbents at the time: Mexican President Vicente Fox Quesada and Health Minister Julio Frenk Mora. Their theory was that if the experiment found positive results, the next government would find it hard to drop the policy.[117] Despite their advantages, cluster trials do have some drawbacks. Some require more complex analysis, and because they randomise at a higher level, they require larger samples – which can be more expensive.

3.6 Stepped-wedge and wait-list designs

USP: A staggered experiment, in which the roll-out of a policy happens in stages and the control group does, in the end, receive the innovation. Helpful for overcoming political or PR difficulties in running a trial, and when we have good reason to believe that people will benefit from an innovation.

It is often not possible to launch the entirety of a new policy or project at once – our example from Mexico's health insurance experiment is one of these cases. There may not be enough funding and resources to do a complete roll-out at the start. Or, there may be cases in which decision-makers want to make sure that the target population of an innovation receives it with as little delay as possible. In these situations, a 'stepped-wedge' trial can provide a feasible approach. As long as it's possible to monitor outcomes in all the areas that will eventually receive the innovation, and there is a willingness to randomly decide which area goes first, a stepped-wedge trial can be used to exploit a staggered roll-out. The design, which inspires the name 'stepped wedge', is shown in **Figure 13**.[118]

Figure 13: A stepped wedge design



Source: Adapted from Hemming et al (2015), 'The stepped wedge cluster randomised trial' in the *British Medical Journal*

One advantage of this type of experiment is that it helps counter the criticism that a randomised trial may mean withholding a beneficial service from those who need it: in this design everybody will, eventually, receive the innovation.[119] Instead of randomising to a simple intervention or control condition, you randomly allocate the time that groups receive an innovation. Trials like this can be used to evaluate interventions during routine

implementation and have a dynamic design, which means that learning early on in the roll-out can be used to inform what's done further down the line.[120] This dynamism does however have some drawbacks: it can make results harder to interpret and analyse, as well as harder to explain to non-experts.

The probation service in the Durham area in the North East of England used a similar design, called a 'wait-list' trial, to test out a new approach to delivering their services. Resource constraints precluded all six probation centres from receiving the new guidance and training at the same time. The fairest approach – and the one that allowed them to learn the most – was to randomly assign the six 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 took part, a robust evaluation of the effects of the new service on reoffending rates could be conducted.[121] One significant drawback of these designs is that they don't allow for the estimation of longer-term effects: because everyone gets the innovation, we can't explore what impact it has in ten or 20 years' time.

In a wait-list design, individuals or groups act as a control group while they wait to receive an innovation, the same principle that underpins the stepped wedge. This approach was used by the UK's Ministry of Housing, Communities and Local Government (MHCLG) in their first ever randomised trial on community integration (**Box 10**). This example shows how flexibility in design can be especially helpful in complex and sensitive policy areas.

Box 10: Building communities through English language support: a wait-list randomised trial

At the moment there is little high-quality evidence on how to help speakers of other languages become more active and at-home in their local communities. In 2016, the UK Ministry of Housing, Communities and Local Government (MHCLG) launched its first ever randomised controlled trial to test whether a Community-Based English Language (CBEL) programme worked to support individuals with low levels of English to improve their language skills and integrate into their wider community.

The trial tested a CBEL course of 66 hours of learning, delivered through 22 classes and 11 conversation clubs to 527 women. It used a 'wait-list' design, which meant that participants were randomly assigned to either receive classes, or to a waiting list. After the trial, everyone on the waiting list was offered the same language support.[122]

The trial showed impressive results. Course participants achieved better test scores in written and spoken English. They also found significant differences on many social integration outcomes, such as new friendships formed with people from other cultures and attending more health appointments.[123]

The trial results helped the Ministry put together new plans for its 2018 Integrated Communities Green Paper, including a network of conversation clubs and a new English language fund.[124] Stephen Aldridge, the Ministry's Director for Analysis and Data, said that the trial increased the Ministry's confidence. *"It has set a new benchmark for the standards we want to apply to determining what works."*[125]

It's worth bearing in mind that different sources can be inconsistent in how they categorise stepped-wedge and wait-list designs – sometimes 'wait-list' is used as a catch-all for both approaches.

3.7 Crossover design

USP: A design in which each '*unit*' (usually a person) receives different innovations in sequence. As well as having a randomised control group, each individual acts as their own control: we can compare the relative impacts of different innovations, against business-as-usual, for a person over time.

Crossover designs originate in medicine, used to study the effects of different treatments on patients. In a crossover design – unlike most other randomised approaches – each *unit* (usually a person) receives more than one treatment, and the relative impacts of different treatment are compared. It's a longitudinal design: each person receives a sequence of different treatments, with repeated measurements taking place across the duration of the study.[126] This might be just two treatments (or a treatment and a control or placebo) – or could be a string of them. As with any randomised trial, participants in the study are randomly allocated into groups. But in this design, each individual receives multiple interventions in a random order. So, participants are randomised to sequences of interventions.[127]

Although most crossover trials to date have been run in medicine, there are plenty of opportunities to run them in other areas of policy and practice. They have already been applied in the wider healthcare and public health sectors – and certainly have uses beyond trialling different drugs. In Tzu Chi University in Taiwan, researchers used the crossover design to the study the impact of listening to music for individuals in high-stress professions like nursing. In one small trial, they measured how two groups of nurses responded to listening to music of their choice for 30 minutes, versus resting in a chair for the same length of time. Music was more effective at reducing the physical symptoms of stress, as well as nurses' own perceptions of their stress levels.[128] In Melbourne, Australia a crossover trial was used to explore how Montessori activities, developed to promote engagement in learning, could form part of a more person-centred approach to care for people living with dementia (Box 11).

Box 11: Can tailored activities improve the wellbeing of people living with dementia?

Over the past ten years many organisations – including Nesta – have argued for a more person-centred approach to health and care.[129] In early 2019, NHS England set out how they will implement a new model of Universal Personalised Care. A more person-centred approach, focused on the needs of individuals and a more holistic understanding of health, is being explored internationally. It often means a big shift in the culture and skills of health and care organisations.[130]

In Melbourne, Australia, a group of healthcare researchers and practitioners wanted to follow up on promising evidence that one-to-one interactions could help reduce some of the behavioural and psychological symptoms of Alzheimer's disease and other forms of dementia. People living with dementia often experience changes in behaviour, and may become agitated, emotional or disengaged. Past evidence suggests that this can happen when their emotional, care or social needs are not being met.

A crossover experiment tested whether personalised one-to-one interaction based on Montessori principles might improve how agitated individuals felt and how engaged in activities they were. The trial took place at nine residential care homes in Melbourne and compared individuals' behaviour during normal care, general one-to-one interaction, and a tailored Montessori task. Montessori tasks aim to promote engagement with learning by creating activities that are tailored to each person's level of ability and areas of strength and weakness. These principles were used to create personalised activities for people in the care home. A crossover design was critical here, because it allowed researchers to separate out the impact of Montessori activities from more general, non-personalised, one-to-one interaction – which is also known to help calm people experiencing these symptoms. The behaviour of participants was observed before, during and after each of the three interventions (normal care, general one-to-one interaction, and a tailored Montessori activity) and their behaviour, emotional state and engagement noted down every minute.

The results of the trial support the value of a person-centred approach to improving individuals' engagement and participation. During both the Montessori activity and the non-personalised one-to-one interaction, agitated behaviour decreased by 50 per cent and 42 per cent respectively. During the Montessori activity, the amount of time spent actively engaged was double that of the non-personalised interaction, and participants showed more positive emotions. The experiment found that general one-to-one social contact can help residents with dementia feel calmer and more settled, but also that tailoring activities to their needs had even greater benefits. This was particularly the case of people who didn't speak English as a first language and had lost fluency. For individuals struggling to communicate in English, there was a significantly larger improvement in their engagement and emotional state during these structured and non-verbal activities.[131]

The crossover design has a number of advantages. First among them is that each participant acts as his or her own control. That is, researchers can compare how a participant responds to an innovation with how they responded to other innovations in the sequence. This reduces the kinds of *bias* that might result from using a control group, if it was in some way different from the intervention group. The crossover design also has some drawbacks. One of these is the 'carry-over effect'. This challenge is intuitive: if an individual is receiving a sequence of three different interventions, it is possible that some of the effects of the earlier intervention might persist into the periods in which later interventions are being tested. Some treatment effects could be delayed, or there may be 'side effects' that confuse later results. Because of this, it is recommended to have a 'wash-out' period when using this design. That is a period between tests, when the innovation – a drug in medical designs – can 'wash out' of the patient's system.[132]

3.8 Multi-site trial

USP: Experiments that increase the evidence base by testing in more than one place. Results may be more generalisable, and these trials are usually accompanied by qualitative research to understand any differences between how things work and are perceived across the different sites.

One common criticism of RCTs is that they only measure one thing, in one time and place. Now, more multi-site trials are being conducted, which have the potential to increase the evidence base on what works, when and where. They are especially valuable because they generate results that tell us more about what's effective for people in different contexts and places. Multi-site trials are conducted in multiple locations and use the same study design so that researchers can directly combine and compare results. Unlike a replication study (which aims to repeat a previous trial, to see if results are the same) multi-site trials are planned and coordinated across sites, often run simultaneously.

One example from the US is an evaluation of a programme that offers comprehensive assistance to families, including nutrition and health support, access to savings, and life coaching. The trial followed 21,000 people over three years, across six countries – a huge achievement. Evaluators found that 'the program generated between US $1.33 and US $4.33 in increased consumption for every dollar spent.'[133] Other trials, like the multi-site trial which built on the 'Rialto experiment' (**Box 12**) uncovered some of the complexities involved in taking promising innovations to scale.

Box 12: The Rialto experiment

The Rialto experiment was the first in a series of randomised trials, run in multiple locations across the world, that investigated whether body-worn cameras could reduce violence and the number of complaints made against police. The Rialto Police Department in California was the first to participate in what became an international trial. There, the technology was a success; researchers found the use-of-force by officers wearing cameras fell by 59 per cent against the previous year and reports against officers dropped by 87 per cent. These results supported the roll-out of the new tech to many US police forces.[134]

The Rialto trial was only the start for Cambridge University's Institute of Criminology. The team replicated the experiment in multiple police forces, from the West Yorkshire force and Northern Ireland's PSNI in the UK, to forces in the US and Uruguay. In total, the experiment was run 30 times. Results from these trials were more complicated. Although they reached the same overall conclusions that complaints against police fell dramatically on average, and there was a significant reduction in the use of force,[135] these patterns did not hold in a small number of locations. To get to the bottom of this, researchers did a meta-analysis (a big statistical analysis of all the results, across the experiments) which covered a population of more than two million and included 2.2 million hours of police work.[136] The analysis revealed that when the protocol for wearing the cameras wasn't followed, they could actually produce negative impacts – the researchers *hypothesised* that this was the case when officers chose to turn off body-worn cameras midway through an interaction or arrest.[137] From these findings, evaluators could recommend how body cameras should (and should not) be used.

A follow-up study also looked at the effects of wearing cameras over time. Did these positive impacts last, or did officers and citizens become desensitised? Taking a longer view, investigators went back to Rialto three years later. They found that the reduction in violence and complaints had continued, in Rialto at least, long after the experiment finished.[138]

3.9 Realist trial

USP: An experiment that aims to create new theory about how and in what contexts an innovation makes an impact, by looking at the underlying mechanisms that generate outcomes.

An important criticism of randomised experiments is that they can sometimes suffer from the 'black box' problem: they are useful for telling us 'what works', but don't tell us much about why, or how effects differ between individuals or settings. This issue is hotly debated, and may not be true of all trials, but it's clear we do need more theory-driven experimentation that aims to shed light on how change happens in the world.[139]

These experiments draw on both qualitative and quantitative analysis not only to find out whether policy innovations work, but also to shed light on who they work best for, and when and where they will do most good. Developed at the London School of Hygiene and Tropical Medicine (LSHTM), a world-leading centre on the evaluation of public health interventions, the realist trial aims to combine the strength of an RCT at making *causal inferences* (or claims about cause-and-effect) with the tradition of realist evaluation, a theory-based approach that explores 'what works, for whom and in what circumstances'.[140]

Realist evaluation looks at the underlying mechanisms that drive change. 'Mechanism' can seem a slightly nebulous term in social science compared with its biological or physical counterpart, but is a word used to describe the processes, actors and activities that cause or create a social outcome.[141] Realist evaluators explore how policies and innovations affect change for participants, in order to develop generalisable theories about what kinds of interventions are effective and why.[142] A realist RCT works by applying some of these principles to an experiment. Instead of focusing only on the primary outcome (the main effect we are trying to achieve with the intervention), it has two aims: first, to assess whether and for whom an innovation is effective, and second to build new and empirically informed theory to understand the effects.[143] Running a realist trial can be complex – and the approach was first laid down in detail in 2012.[144] It starts with developing a detailed theory of change or logic model, and then refining and testing this during the experiment. A theory of change is a tool that sets out what an innovation aims to achieve, and how it plans to achieve it. Crucially, the theories of change used in realist evaluation take the form of what's called 'CMO' configurations – or 'context-mechanism-outcome'. These propose how the context of an innovation interacts with its key mechanisms, to generate outcomes. Realist trials draw on qualitative research to refine the theory of change and then test this using data on outcome measures. With a realist RCT, the experiment is used to help develop theory. It aims to test some of the mechanisms through which the innovation is *hypothesised* to work, and uses this learning to strengthen the theory we had about how to make change before the trial and refine the innovation for those who might use it in future.

The first realist trial was developed for an education innovation called Learning Together, which uses restorative practice to bring together pupils involved in bullying, conflict or misbehaviour (**Box 13**). Based on theory about human relationships, the experiment allowed researchers to develop and test theory about a more holistic approach to building children's sense of belonging and relationships in school, and the consequences of this for rates of bullying, as well as various health outcomes.

Box 13: Learning together: Lessons in theory-driven experimentation

Learning Together is a project run in the South East of England, which has brought together pupils involved in bullying or conflict to appreciate the harms of misbehaviour. This approach, called restorative practice, originates in criminal justice. Learning Together was developed by a team at the London School of Hygiene and Tropical Medicine, but was informed by ideas from the Gatehouse Project, a trial in Australia which took a new approach to developing children's social and emotional skills.[145] Rather than over-burdening schools with multiple interventions, this project aimed to provide schools and educators with one coherent programme.

Over the years, much rich theory has been developed about how human relationships function, and how this relates to the organisation of schools.[146] The team drew on systematic reviews, which brought together all the available research on whole school approaches to reducing bullying and improving health and emotional outcomes. From this evidence, researchers developed a detailed theory of change to plan the innovation.[147] This *hypothesised* that students who didn't have good relationships at schools, and a sense of belonging and participation, were more likely to engage in bullying and other risky behaviours. To combat this, the theory argued, schools can improve relationships between staff and students, train staff in restorative practice, and develop a social and emotional curriculum.[148]

Before and during the trial, mixed methods research explored the theory of change, which described the mechanisms that should be triggered by the innovation, as well as how these mechanisms interacted with school context to generate outcomes that might vary between schools and students. This helped the team to understand which aspects of evidence were most relevant for the schools they were working with, so they could refine their model. The trial examined not only overall outcomes for students, but also the mechanisms that seemed important to achieving them. They did this by paying special attention to what are called the 'mediators' – mechanisms that affect outcomes – as well as 'moderators', the factors that explain how effects vary between schools and students. The study found hugely positive results. Learning Together helped tackle bullying, improved mental health, and lowered smoking, alcohol and drug consumption.[149] In addition, the team were able to look at how these changes may have resulted in improvements to the school environment, which strengthened relationships and helped students disengage with disruptive peers.[150] This new learning, that builds on theory we had before, is now being fed back into the programme's theory of change, before Learning Together is offered to other schools.

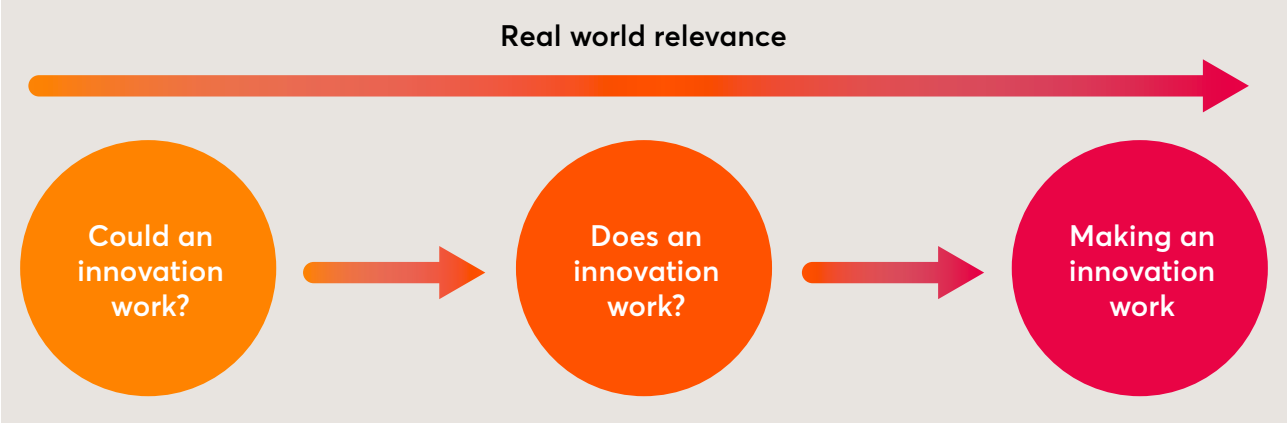
Realist trials have created some controversy – and a long debate in the academic journals Social *Science & Medicine* and *Trials* – between advocates of the new approach and those who believe it's not possible to do experimental tests on realist theory. Realist trials generally need to be larger in order to assess how effects vary between schools and students. Their supporters argue that the extra effort could be worth it, as a better understanding of how interventions work could save the public money in the long term and increase the quality of the innovation in its final form.

3.10 Hybrid trial

USP: An experiment with two goals: to test an innovation and find out how it should be adopted in everyday practice.

Hybrid trials have two aims: to test the effectiveness of a new idea, and also how it should be implemented in the real world, incorporated into the everyday practice of an organisation and its staff.[151] This approach originated in the work of health researchers in America who wanted to speed up the adoption of promising ideas from research into practice. They were interested in a new discipline that looks at how evidence turns in to action, called implementation science. Implementation science studies how ideas that are backed by research can travel through the pipeline from science to routine care. Researchers propose that promising ideas from basic research are first trialled under ideal conditions (like the 'efficacy' trials we discussed in **Box 4** earlier), and then conditions that aim to approximate the real world (a more 'effectiveness' trial). Finally, implementation science argues, research should focus on the concrete strategies for the use of innovations by individuals, teams, and whole organisations. This 'pipeline' is shown in **Figure 14**. As we move through it, experiments become more relevant to the real world.[152]

Figure 14: The 'pipeline' of experimental research, according to Implementation Science



Source: Adapted from Brown et al, 2017

Hybrid design aims to tackle these last two phases – whether an intervention works, and how to make it work in practice – at the same time. This aims to speed up the translation of research and create new evidence on what makes promising ideas work on the ground, including insights that are more useful for professionals and leaders tasked with enacting policy.[153]

A key difference between a hybrid trial and a normal RCT is an explicit focus on measures of organisational processes and outputs, as well as on outcomes for the beneficiaries of an innovation. They ask: which strategies make a policy idea more efficient? Or have better quality of delivery?[154] A group of researchers working for a healthcare initiative for veterans in the US first defined three types of hybrid designs in a 2012 article.

Type one looks most like a basic RCT, where an impact study is combined with a process evaluation. This process evaluation gathers information on how an innovation is being delivered in practice, during the trial. An example of this is the CALM (Coordinated Anxiety Learning and Management) study, that tested the effectiveness of the CALM innovation to help people with anxiety disorders. During this RCT, a team conducted qualitative research with multiple stakeholders to understand what the barriers and facilities to CALM's success were, what challenges practitioners faced, and how to improve implementation in future, for the practicalities of the clinic or hospital.[155]

Type two is a simultaneous test of an innovation and the way it's being delivered in routine practice. So, the trial measures outcomes both about the innovation's effectiveness, and the effectiveness of the strategy that's used to deliver it. This can be a bit mind-boggling, but an example is testing a new drug, while simultaneously testing the efficiency of the system through which people are offered it. These have been conducted as large randomised trials, but running them can be challenging, and the design does have important downsides.[156]

Type three reverses the emphasis of a traditional trial, testing competing strategies for implementation (how an innovation is delivered) to see which works better to support quality and efficiency of delivery. As a secondary priority, these trials also gather information on impact. This was the case in the Community Youth Development Study, that trialled how community leaders could make decisions about young people's drug use in 24 randomly selected communities across the US. The Community Youth Development Study tested an approach to planning prevention initiatives among communities, building the capacity of local leaders. The study found this increased the use of science-based prevention techniques, resulting in community-wide improvements in youth development.[157] Now, a follow-up study is looking at the long-term effects on young adult substance use, violence and crime.[158]

Because a hybrid trial is measuring several things at once, it is typically more costly than a simpler form of trial. Accordingly, a hybrid trial is most useful when either we have good evidence that the idea being trialled will genuinely benefit recipients, or we are facing significant time pressure such that we cannot afford to wait for the results of one evaluation before starting the other. (In such a time-limited scenario, it may be wise to test multiple treatments in this way if we have enough trial participants.) Unless we get implementation right, we risk many innovations stumbling at the final hurdle and a hybrid trial can protect against this possibility.

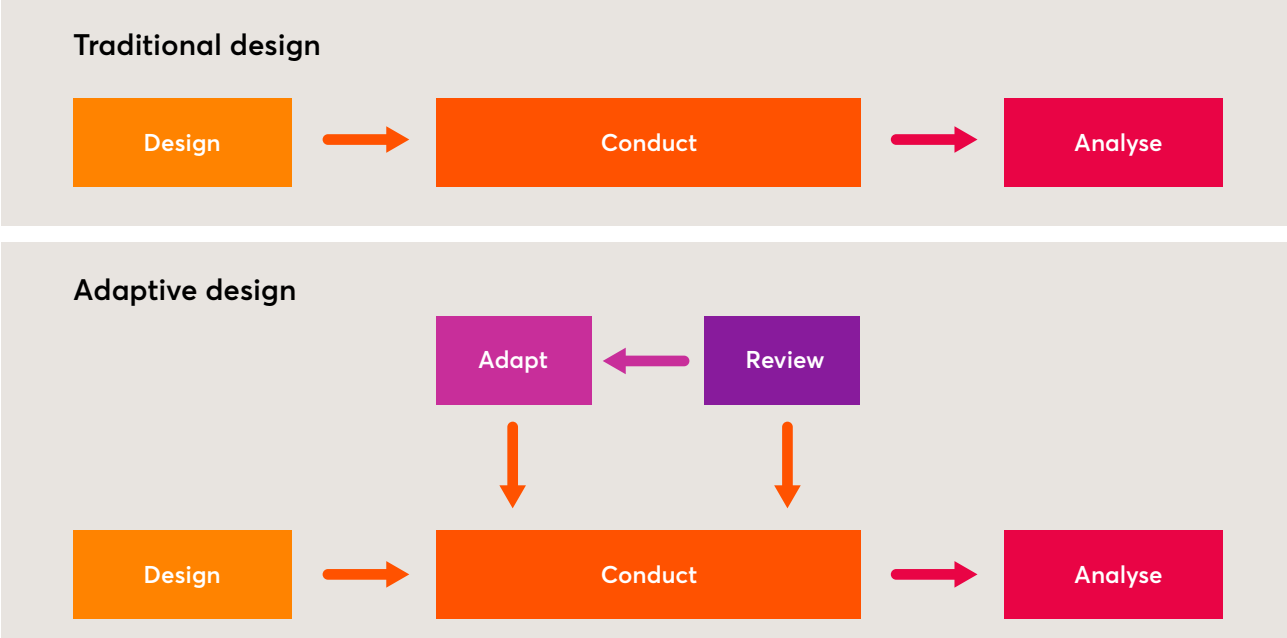
3.11 Adaptive trial

USP: An adaptive trial 'plans to be flexible'. Early results modify the course of the experiment, according to pre-planned criteria, making it possible to test multiple interventions or intervention variants more effectively.

Most randomised trials are run in three phases. First, the trial is designed. Second, it's conducted in line with the specified design. Finally, data is analysed. Adaptive trials work differently. They aim to make randomised experiments more effective by using early results to modify the course of the experiment. These modifications must be planned in advance and serve to better test a particular *hypothesis* – so an adaptive trial is not the same as making ad-hoc changes to an experiment once it's up and running. Some researchers think of this approach as 'planning to be flexible':[159] because they have built-in flexibility, adaptive trials have the potential to be more efficient and informative.[160]

There are different motivations to build an adaptive experiment. Common ones include the ability to drop interventions that don't look promising early, to more quickly identify people who benefit from an innovation, or to stop a trial early if it's unsuccessful or not going to plan. One type of adaptive trial used in healthcare, called MAMS (multi-arm multi-stage), tests multiple treatments at the same time (see Section **3.2**) and has the option to drop losing treatments and identify winning ones early. All adaptive designs use interim data analysis to modify the trial, without compromising its validity (ability to answer the research question) or its integrity (its process).[161]

Figure 15: Adaptive trial design



Source: Adapted from Pallman et al, 2018

Despite being around for more than 25 years, adaptive trials are rare outside of health. They are complex to run, requiring specialist technical and practical experience. But there is growing potential to use these intelligent designs in social policy; some applications that use machine learning are already being used by tech companies and businesses (see **Box 6** on 'Multi-armed bandits').

The adaptive trial isn't a 'free lunch'. Besides the greater expertise required to operate an adaptive trial, they generally require more stringent statistical standards, which in practice necessitates a larger sample size. A trial cannot be 'converted' to an adaptive trial once it has started, either; this is a process known as 'p-hacking' and the results from such trials are untrustworthy. Despite these drawbacks, if you have sample size and expertise to spare, an adaptive trial could allow you to test multiple treatments or treatment variants more effectively. As with any design, its utility would depend on your aims.

Non-randomised and quasi-experimental designs (QEDs)

USP: Non-randomised experiments and QEDs use inventive designs and statistical techniques to create a comparison group to estimate the effects of an innovation. They are particularly useful when randomisation isn't possible or desirable, or when a new policy is already in place.

When a decision has been made to implement a new idea or policy without a randomised trial first, or when policymakers want to make sure that everyone who is eligible for a new scheme receives it immediately, a quasi-experimental design can serve in place of a trial in some circumstances. It is recommended that QEDs include research on the individuals and communities targeted by the innovation as well as the statistical evaluation – and some evaluators argue this is a key strength of the designs.[162]

Non-randomised designs use existing data and exploit natural variation in policy and between populations to create comparison groups. In this section, some of the examples we draw on are '*natural experiments*' (see **Annex A** for a definition).[vii] Because they don't use a randomly allocated control group, quasi-experimental designs typically have to assume that a certain group of individuals is comparable to the intervention group: that is, that they would have had similar outcomes if they had both received the intervention (or if neither had). Sometimes this assumption can be tested and sometimes it can't. Using a non-random comparison group means we need to think through two key challenges. The first is *selection bias* (also known as *confounding*). The challenge relates to how a group is selected to receive an innovation – and whether this in itself makes them different to any comparison group. We will see some examples of how this might happen below. The second challenge is how to take account of population differences that may not always be observable. Unobserved factors are things about individuals or groups that we don't know about or can't measure, but still might have an effect on the outcome we are trying to improve.[viii] This is *confounding*: an effect which we think is caused by the intervention might actually be caused, or partly caused, by whatever was different between the intervention and comparison groups. *Confounders* distort the apparent relationship between an innovation and outcomes; if we don't take them into account, our estimates are less credible.[163]

Confounders are a threat to any experiment, randomised or not. But in a randomised trial, we can usually expect that the randomisation has produced two comparable experimental groups. Consequently, none of the methods in this section have the strength of a random experiment in helping us make confident claims about cause-and-effect. It's worth keeping in mind that there are some disagreements about the value of QEDs, but there is substantial evidence that, in policy areas where previous outcomes are good predictors of future outcomes, QEDs can be a strong design choice. This is the case, for example, in some areas of education. QEDs can often replicate the results from randomised experiments, although this is not always the case.[164] Despite some imperfections, they remain a valuable part of our arsenal for understanding what makes effective policy. There are some situations in which a randomised experiment is just not possible: when investigating the effects of changing prison sentencing a randomised experiment is likely to be unethical; or if a policy has already been enacted it may simply be impractical to use an RCT.[165] In these situations, we must have a range of useful methods at our disposal. Quasi-experimental designs have other benefits too; they can for example be less costly since they usually rely on existing data rather than having to collect their own.[166] One kind of QED, called 'instrumental variables', we don't cover in this inventory because opportunities to use it tend to be few and far between.

QEDs and non-random experiments first gained prominence in social science research, but are now explored in health, and political science too. They have been used to test innovations in public health, economics, criminal justice, education and many other policy areas.[167] Because they use statistics and tend to be heavy on jargon (blame the economists) they can be tough for non-experts to understand. But these approaches play a fascinating role in the history of experimental decision-making.

3.12 Regression discontinuity design

USP: An RDD can be used where an innovation is delivered to those who fall on one side of an arbitrary cut-off (such as being eligible for a certain benefit if your income is below a particular level). It compares those just above the cut-off with those just below it to estimate the effect of the innovation.

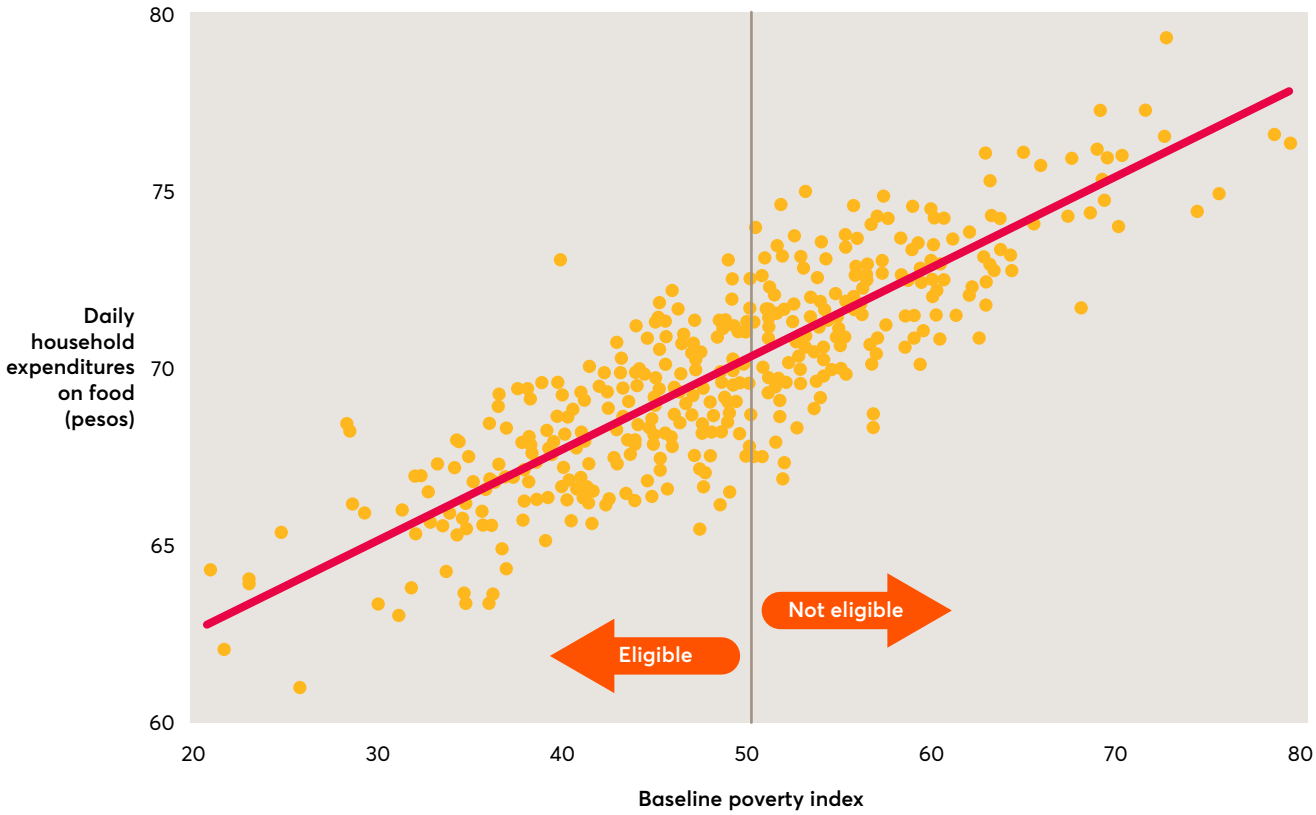
Regression discontinuity design (RDD) was invented in the 1960s by Donald Campbell, the proponent of an 'experimenting society' after whom the international Campbell Collaboration research network is named.[iv] In a paper published in 1963 Campbell and his colleague Julian Stanley argued that regression discontinuity design was able to mimic a randomised experiment by exploiting the way in which social policies or interventions are allocated to individuals who receive them. At Northwestern University in Illinois, where Campbell taught, a group of students tested and refined the design. In one real-world test, they examined the effect that being put on the Dean's List – an accolade for getting good marks – had on their fellow students' performance the following term. At that time, students were put on the Dean's List if their GPA (grade average) was above 3.5. Students in Campbell's group tested whether being put on the list affected performance by comparing those who narrowly got on the list, with those who just missed out.[168] In doing so, they used the idea of a 'cut-off' point – a numerical threshold, fixed arbitrarily to determine who gets what – as a research design principle. Donald Campbell lost interest in RDD over the years, but the design was reinvented by students of psychology, education, statistics and economics, where it was given several different names, such as the 'cut-off based design'.[169]

The central idea to the RDD is that individuals just above and just below the threshold would have had similar outcomes if the treatment didn't exist. In the example above, a student with a GPA of 3.49 is not materially different to a student with a GPA of 3.51. Accordingly, any difference in their outcomes can be ascribed to being on the Dean's List.[170] Most programmes or innovations have some kind of entry criteria that determines who is eligible to participate and who isn't. This might be gender, location, test scores, age or income level. When this entry criteria is determined by a numerical scale with a cut-off point – so age, income or test scores from our list above – we can use an RDD to look closely at the innovation's success. An RDD creates a comparison group from those people who fall just outside of the cut-off point for the policy in question. This works because there should be no systematic difference between these people and those just inside the policy boundary; for example, in relation to the factors that are relevant to the Dean's List (i.e. grades). This is what led Campbell and Stanley to argue in 1963 that RDD mimics a randomised experiment around the cut-off point of a policy or project.[171]

RDD can only be used where a programme or innovation has a numerical threshold or cut-off; this must be continuous, like a quantified index of some kind, on which the population of interest are ordered or ranked.[172] It is also important that the individuals can't manipulate their score, or that the intervention doesn't incentivise them to do so. In the GPA example, the students were trying to maximise their GPA anyway, but if there was a benefit being offered to students with a low GPA then that could be problematic to evaluate with an RDD. This is because students might deliberately score a low GPA so they could receive the benefit.

The second condition for an RDD to be effective is that the cut-off point must be arbitrary – that is, a point chosen for convenience, that could just as easily be a little higher or lower. The following graphs from the World Bank (**Figure 16** and **Figure 17**) illustrate this with the example of an anti-poverty policy called a cash transfer scheme.[173] In the first graph, the vertical axis is the household expenditure on food, while the horizonal axis represents total household income, calculated as a number on a 'poverty index'. Point 50 on the poverty index represents the 'cut-off' for this scheme. Those families below 50 on the poverty index receive regular cash transfers to spend on household goods, while those above it do not. In the first graph – and as we'd expect – we see household spending slowly rise as family income rises.

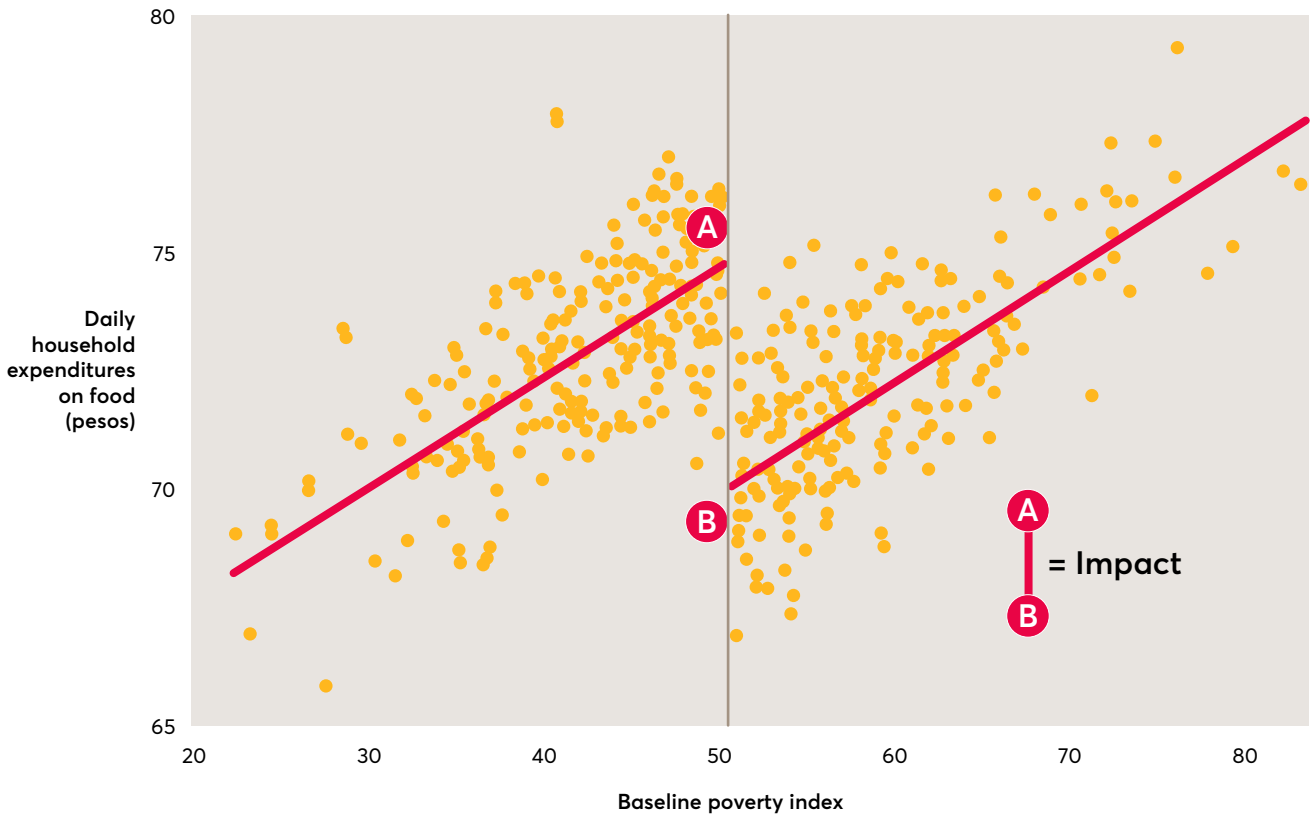
Figure 16: Evaluating a Cash Transfer Progamme with a RDD. Household expenditure in relation to poverty (pre intervention)



Source: World Bank, 2011

In the second graph we see what happens after the innovation is introduced. Now, there is a break in the trend. Families who are poorer seem to be spending comparatively more on food. The distance shown on the graph at the cut-off point shows the impact of the policy: it has been successful in helping poorer families purchase food.

Figure 17: Evaluating a Cash Transfer Progamme with a RDD. Household expenditure in relation to poverty (post intervention)



Source: World Bank, 2011

This is where regression discontinuity design gets its name: when a policy has an impact – either positive or negative – it produces a 'discontinuity', or a disconnect, between outcomes for those who received the policy intervention and those who didn't, if we look closely at those people near the policy threshold.

The final condition is that the threshold is only the cut-off for the policy you are trying to evaluate. If other policies use the same cut-off, then you would be evaluating the effect of all of them together. For example, free school meals are offered in England if parental income is below a certain threshold. However, at the time of writing, the same threshold also entitles the parent to Universal Credit. If we used RDD we would not be evaluating the effect of free school meals; we would be evaluating the combined effect of free school meals and Universal Credit (and anything else which uses the same income threshold).

Day-to-day policymaking provides plenty of opportunities to use this design, as many schemes are allocated using numerical scales and criteria – even how people are sentenced to prison, as researchers who conducted an RDD to explore the impact of imprisonment on reoffending discovered.[174] An RDD is a great design choice in a case like this, where a randomised experiment would not be possible to conduct. But it does have a key drawback. Because it's focused on people either side of a policy threshold, the results aren't usually generalisable to everyone in the study or in the broader population. They are however still useful for many policy-relevant decisions, like whether the entry criteria for a policy benefit should be changed, or if a programme should be cut.[175] RDD does not provide us with answers about impact on the whole population.[176] Because they only consider a small subset of the wider population, RDD usually requires a larger sample size than a randomised design. This is not normally a problem for programmes implemented on a national scale, if nationwide administrative data is available, but may be a significant issue on smaller scales.

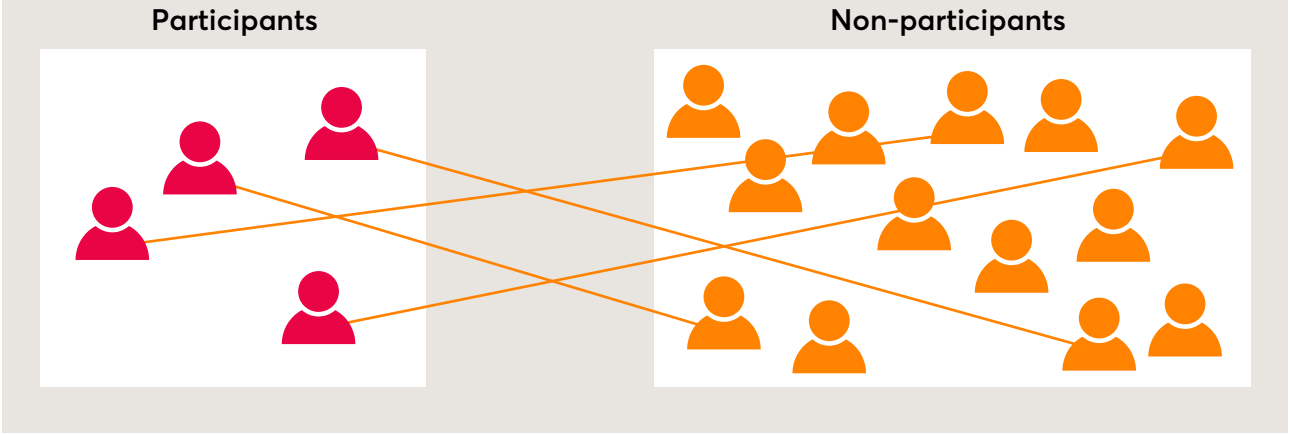
Researchers use statistical tests and models to explore complex relationships between the comparison and intervention group, and account for potential sources of *bias*. In a 2002 book on making *causal inferences* (claims about cause and effect) Prof William Shadish and colleagues published a helpful table of different ways to strengthen the basic RDD.[177] For those newer to evaluation, the World Bank's *Impact Evaluation in Practice* book is free online and an excellent guide. For a more technical audience, Lee and Lemieux (2010) provide a good guide.[178]

3.13 Matching

USP: Matching is useful when there's lots of data available but it's not possible to randomise – it creates a comparison group by 'matching' people who receive an innovation with similar people who don't. It's best if people had little or no control over whether they received the intervention.

The main principle of a matched experiment is to create a comparison group by matching '*units*' – those might be individuals, households, populations or areas – who receive an innovation, to apparently similar '*units*' who do not. We judge the similarity of *units* based on observed data, such as the age, gender and income of individuals or the inspector's rating, grade average and number of pupils in a school.

Figure 18: Matching



Source: Behavioural Insights Team

The success of a matching approach turns on how comparable a matched sample really is. Suppose we want to evaluate the effect of a testicular cancer screening programme, which anyone is eligible for but has been marketed to the over-55s, and requires patients to proactively book an appointment with their doctor. Bob the 56-year-old man has received the intervention and is matched with Rob the 56-year-old man who has not. Bob and Rob are comparable on the data that we have, but the knowledge that Bob proactively booked an appointment and Rob did not has changed things: Bob might be more health conscious than Rob (and will probably have better outcomes which we will mistakenly attribute to the screening programme), or Bob might know that he is at particular risk of testicular cancer because a relative had it (in which case he will probably expect a worse outcome and we will mistakenly conclude the screening programme is harmful). If the patients had less active control over whether they were screened (for example, if certain GPs offered the service to everyone in their practice but other GPs didn't), then matching would be more appropriate.

One matching technique that has gained popularity in recent years is propensity score matching (PSM), coined by Professors Paul Rosenbaum and Donald Rubin in a 1983 article. Rather than matching directly on characteristics such as age or school size, PSM first uses this data to predict how likely a *unit* is to have received the intervention. For example, a school with a 'good' rating and 1,000 pupils might have a 10 per cent chance of receiving an improvement programme, whereas a school with a 'requires improvement' rating and 250 pupils might have a 75 per cent chance. Once these 'propensity scores'[179] (from which the method takes its name) have been calculated, treated *units* are matched with untreated *units* with a similar propensity score to create a comparison group. This avoids a problem which often arises, that once you have several pieces of data relating to a *unit* (person, household, population) it is hard to find a comparator that matches all of them.[x] PSM only matches on the measure that matters: the chance to have been treated.

Once a propensity score has been calculated for everyone receiving an innovation, and everyone in the potential pool of comparison, 'matching algorithms' pair up beneficiaries with non-beneficiaries. The use of an algorithm helps to produce a close match.[180] Once all the matches are selected, researchers can compare average outcomes for the two groups.[181] They will also run tests to make sure the comparison group selected is similar enough.

There are plenty of pros and cons to a matching design. One strength is that because it involves bringing together lots of data, it means conducting thorough background research on all the factors affecting outcomes. It will also draw on theory of how the innovation works, in order to explore which factors are most critical. All this research means that PSM can help answer some detailed and policy-relevant questions about who a policy innovation might benefit the most.[182]

The downside of this is that PSM requires very good data, and lots of it – usually high-quality administrative data, often linked to other sources, or perhaps a large survey. The more holes there are in the data, the weaker the matches will be, and the less reliable the results. These strengths and challenges come out in the example below, which draws learning from one part of the evaluation of Sure Start (**Box 14**), an early intervention programme that has aimed to improve the lives of children and their families in England since 1999.

Box 14: Studying the effects of early intervention over time

Sure Start was introduced in 1999 to improve life chances for young children growing up in highly disadvantaged areas. It took a new approach to early intervention by trying to work in a more holistic way with families and communities. It was one of the first – and remains one of the largest - 'area-based' policy programmes in the UK.[183] In the early years of the project local areas had lots of control over how services were delivered, and local Sure Start teams developed community-specific solutions. This changed in 2005 and 2006 when local authorities began running the programme as Sure Start Children's Centres. By 2009-10 Sure Start accounted for £1.8 billion in public spending.[184]

An evaluation, started in 2001, aimed to study what the effects for children and families were over time. It looked at the impact on 500,000 children living in 150 communities as they grew up. Here, we look at how the evaluation team used propensity score matching (PSM) to find out what effects the programme had for children aged seven, over two years after their last contact with the programme.

Sure Start was rolled out quickly, and a decision was taken by the government not to run a randomised trial at the time. But evaluators were able to draw on a longitudinal dataset called the Millennium Birth Cohort Study (MCS) to

create a comparison group. MCS has followed the lives of around 19,000 young people born across England, Scotland, Wales and Northern Ireland in 2000-2001.[185] For Sure Start evaluators, the MCS provided a vital source of data which enabled them to use a PSM design. They identified comparison areas and were able to shed light on how Sure Start had impacted families over years. They identified 172 comparison areas, including 1,879 children.

Despite the data available, the study faced methodological challenges. There was a two-year gap between the Sure Start data and the Millennium Cohort Study data, as the two hadn't started at the same time. In addition, MCS did not include as many economically disadvantaged families as Sure Start. These factors meant that evaluators had to do additional statistical work to reduce *bias* in the study, and the *bias* could not be eliminated entirely.[186]

Despite these challenges, they found some valuable results. Sure Start had beneficial effects on family function and maternal wellbeing which persisted to age seven. Mothers in Sure Start areas reported engaging in less harsh discipline and providing a more stimulating home learning environment for their children. Importantly, the study was also able to explore some of the impact on more disadvantaged groups. In the past, evidence had suggested that the most disadvantaged families – with single parents, or no working family members – experienced some adverse effects as a result of the intervention. This time, the researchers found positive rather than negative impacts. Lone parent and workless households reported better life satisfaction than the comparison group.[187]

Teams providing Sure Start had learnt from past negative results and were engaging more productively with more vulnerable groups.[188] These positive effects had persisted years after the end of early years care, although the evaluation found no consistent impacts on child educational development, social or behavioural outcomes. Despite the challenges, evaluation has provided a vital source of learning for the programme over the years, which in June of 2019 was found to have major health benefits for some of the poorest children in England, reducing hospital admissions by up to 19 per cent by the time children are 11.[189]

As the evaluation of Sure Start shows, PSM is most useful when we have lots of information; it requires an evaluation team with thorough knowledge of the sector and context, as well as lots of data and a large sample size. It relies on there being a wide enough range of possible comparisons available to create overlap between people receiving an innovation and those who don't; this overlap is technically called 'common support'.[190] Importantly, PSM assumes that there are no unobserved differences between the comparison group and those receiving an innovation, like a behavioural characteristic that makes someone more or less likely to respond well to a policy idea. This cannot be measured but it should not be ignored, and it introduces an important potential source of *bias*.[191] But, when done well and used in an appropriate situation, PSM can provide rich information about the participants of a programme, and how impacts might be different across population sub-groups, though it can't give us as much certainty as a randomised trial – and getting it right is often a challenge.[192]

3.14 Difference-in-difference

USP: DiD comes in handy when a policy is introduced in one area or region, but not in another place that's similar or has comparable trends, and is on a similar trajectory pre-intervention. It compares the change after intervention in the treated area with the change after intervention in the comparison area (hence 'difference-in-difference').

In 1855 John Snow, a doctor working in Soho, London, published research that proved that cholera was a water-borne disease. Before that, the medical community believed that the disease – spread by sewage dumped into London's waterways – was transmitted through 'miasma', or 'bad air'. John Snow became the founder of modern epidemiology, the discipline of how health and disease work at scale – and his breakthrough study of cholera also invented the difference-in-difference design.[193]

Difference-in-difference (DiD) tracks what happens to a group who receive an innovation over time, before and after an innovation is introduced. It also tracks changes for an identified comparison group. It then compares trends for both groups and calculates the relative change in outcomes between the two, to arrive at an estimate of impact.

DiD relies on taking thorough (and ideally multiple) baseline (or pre-test) measurements about those who receive an innovation and those who don't, to understand what is happening for both groups before the innovation is introduced. Measurements are then taken after the innovation is introduced – or, in some designs, at several different points – to see what has changed. The effect of the intervention is then estimated as the difference between the change in the 'treated' group and the change in the 'untreated' group (hence the name 'difference-in-difference').

DiD is only useful when the experiment and comparison group have historically followed the same trends over time and would be expected to do so in future. This can be partly verified: we can check that the trends were similar historically but not whether they would have stayed that way if no intervention happened. It is also important to understand that we only need the trends to be the same, not the overall level (so an outcome like pupil attainment doesn't need to be the same in both, just the direction it is going in over time). A DiD can be used between two settings that have radically different outcome levels overall, as long as when one setting experiences an increase or decrease in the outcome measure, the other one does as well by a similar amount.

A DiD study that explored the impact of abolishing school league tables in Wales provides a strong example of this (**Box 15**). Wales and England have historically similar education systems that have followed the same patterns in the past, even though Welsh and English outcomes are not at the same level, they tend to both rise or fall together.

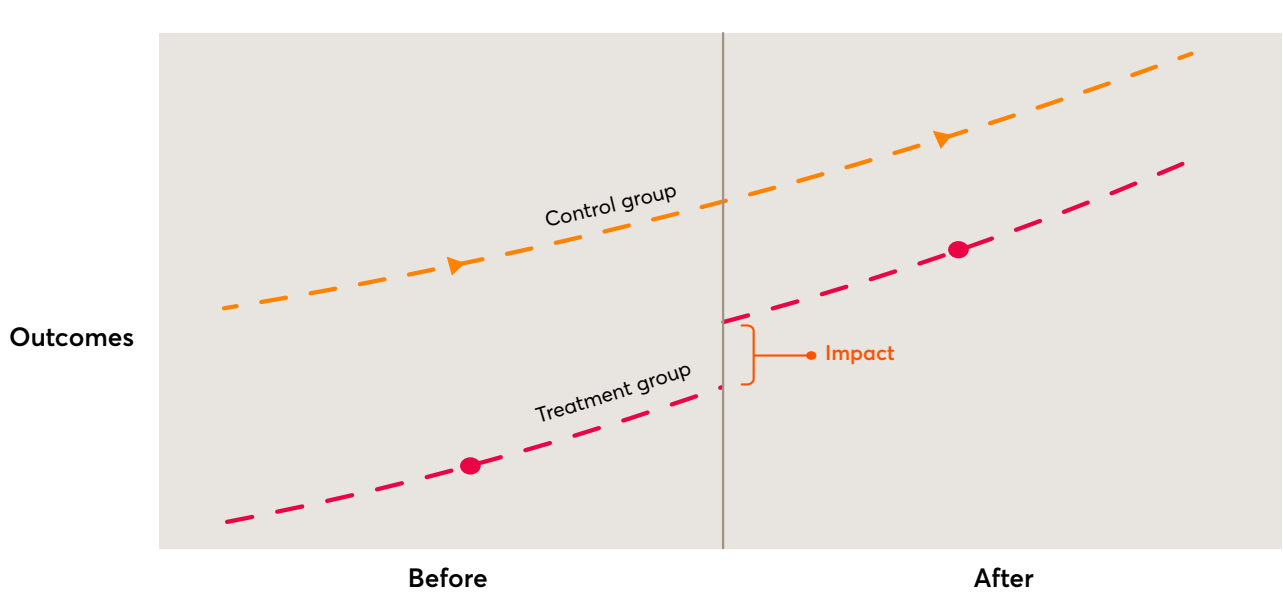
Box 15: Do school performance tables raise educational standards? The difference-in-difference between Welsh and English education policy

Between 1992 and 2001, secondary school performance tables were published annually in both England and Wales. After the devolution of power on education policy, the Welsh Assembly took the decision to abolish the publication of these tables in 2001. Other aspects of the education systems remained the same: both followed a common curriculum, kept the same exam processes in place, and had the same school inspection system. In almost all aspects, 'the education systems of the two countries were practically identical until devolution of power' – this was the only policy change differentially affecting the two school systems.[194]

This similar set-up – and historically parallel trends – allowed researchers to investigate the effects of the change, and evaluate the effect of publishing league tables, by using a DiD design. This meant investigating the changing patterns in school performance in the two nations before and after the policy change. They also had a third point of comparison: primary schools, where no change had been made (the third dimension of whether a school is primary or secondary gives rise to what researchers call a 'triple difference'). The researchers *hypothesised* that removing the accountability provided by school league tables could have a negative effect on outcomes – and they were right. They found that doing away with school league tables reduced the effectiveness of Welsh schools. On average, the percentage of students receiving at least five good GCSE grades fell by 3.4 per cent per school, though the top-performing quarter of schools was unaffected.[195]

When the comparison group is chosen well, trends between the two groups should follow two roughly parallel lines – with school grades consistently rising, for example. After the introduction of the policy, impact can be seen on the graph if outcomes for the experiment group start to deviate from the trend.

Figure 19: A difference-in-difference estimation of impact



Source: Adapted from Columbia University, *Population Health Methods*, as of 2019

Devolution, regional, local or phased policy changes all provide good opportunities to use the difference-in-difference design.[196] It requires ample data, but is a valuable option for exploiting '*natural experiments*', when geographically specific policy innovations are implemented and we want to understand what impacts these have had. DiD rests on two key design assumptions. The first is the assumption that the trends are parallel, something we can easily see in **Figure 19**. 'Parallel trends' assumes that trends for participants and non-participants would have been equal over time if there were no intervention. This includes any trends we aren't able to observe or measure.

Researchers can use statistical tests to try to investigate the parallel trends assumption as long as historical data is available. (If it isn't, then DiD is a very risky choice as it relies on untestable assumptions). Generally, it is wise to compare the trends with other settings where we would also expect them to be parallel (other regions which are not part of the study, for example). There may be unobserved factors that could easily trip up a DiD design. A nice example is a study looking at the impact of a vaccination policy. We could ask whether there are things that might affect vaccination rates across places and time and that we don't have good measures for – like parental attitudes, which are strongly influenced by peer groups and can change in one location but not another. In this case, where we have a clear example of an important factor we cannot measure and which might not obey the parallel trends assumption, DiD will not be a good design choice.[197]

The other design assumption is called 'common shocks'. This means that, for a DiD analysis to hold true, any unrelated events (or shocks) that occur within the period we are studying are expected to affect both groups equally. For example, if the length of the school day were to change across the UK, we'd need to assume that students in England and Wales reacted the same way. While researchers can take measures to reduce the threat of *bias* from parallel trends, we can't test whether or not we are right in assuming that both groups will react the same to common shocks and this limits the robustness of DiD designs.[198] The impact of an unrelated event can introduce what academics call 'history bias', which may distort findings.[199]

For more complex policy innovations or research studies that involve multiple groups receiving an innovation, more nuanced DiD designs do exist. The main one is a 'difference-in-difference-in-difference' (tongue twister, we know), where another group is added to the design. In the evaluation of school policy in Wales, researchers introduced a point of comparison, primary schools, making the study a 'triple difference', or DDD for short.

A DiD design cannot be used if other major changes have occurred in the same time period. And, it requires that good data is available across the whole time period in question, which can raise some immediate practical challenges.[200] While DiDs are an imperfect design, they can be a pragmatic option for understanding the impacts of a policy that's already in train.[201] If an RDD is available then it is usually preferable to a DiD, since it makes far fewer assumptions and they can generally be more rigorously tested.

3.15 Synthetic control

USP: A data-driven approach useful to understand policy changes that have already taken place, by comparing many individuals, populations and places over time.

Synthetic control is a relatively recent and promising approach developed by two researchers from the Basque Country, who together wanted to understand the impact of terrorism on economic growth in the País Vasco, a contested region of Spain. Their study worked by creating a comparison for the Basque Country from a collection of other regions in Spain. It used available data collected from different places and times to construct this 'synthetic' comparison region, based on a combination of statistical averages.[202] This is the 'synthetic control': a synthesised comparison made up from places or people – called the 'donor pool' – who may not individually be comparable to the intervention group, but taken together display similar characteristics. For example, if an intervention region has 30 per cent of its population from a particular ethnicity, then it could be approximated by taking two-thirds of an untreated region with 40 per cent from that ethnicity, and one-third from a region with 10 per cent.

To do this in practice when there may be multiple dimensions and metrics to balance, the researchers invented a novel statistical approach. It was popularised in 2010 when a second study was published, accompanied by a free, open-access software programme called Synth that helps anyone use the synthetic control method.[203] In a study that investigated California's anti-smoking law Proposition 99, the researchers used synthetic control to argue that the legislation dramatically reduced tobacco sales (**Box 16**). To do this, it looked at relevant variables – like data about smoking behaviour – and studied how it changed over time in California and many possible comparison states. Relevant factors and outcomes in these states were averaged, and averages given more or less statistical weight in the analysis. When combined, the result is a 'synthetic California' that simulates what the states' smoking outcomes would have been, if Proposition 99 had not been passed.[204]

Box 16: Using synthetic control to measure the success of California's anti-smoking laws

In 1988 California passed Proposition 99, a law which increased tax on cigarettes by 25 cents a pack. These taxes produced over $100 million in revenue, which was earmarked for spending on health and anti-smoking education budgets, as well as media campaigns and clean-air policy. But researchers wanted to find out: did it work to reduce smoking? No state in the US was directly comparable to California. Trends in smoking were varied; and no individual state brought together all the factors needed to create a good enough comparison. Synthetic control allowed them to find the attributes and patterns that matched California's in a range of states and bring them together to create a model comparison state.

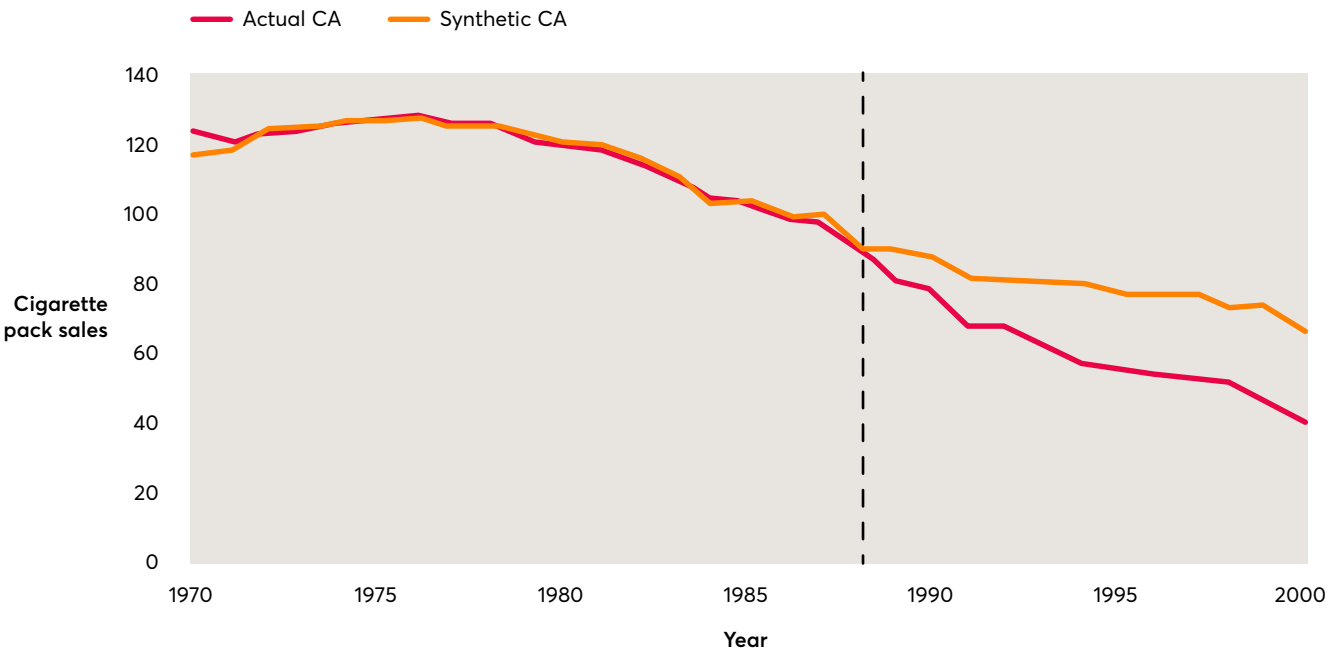
They started by identifying what factors would predict changes in smoking behaviour before the policy change was introduced: the price of cigarettes, the age of the population, how much beer they consumed, and what income levels were for example.[205] They looked at how those factors had changed in the past, to get a sense of how trends went over time. Then they identified comparable states in the period before the innovation, going back as far as 1970. Using synthetic control, Alberto Abadie and his colleagues estimated that, by 2000, sales of cigarettes had dropped by 26 packs per person due to the legislation. This was a much bigger impact than earlier estimates had suggested.[206]

For this method to be useful, no similar policy measures can have been introduced in the relevant time period which would have an effect on the innovation or the outcomes of interest, in this case smoking behaviour. Synthetic control shares some similarities with difference-in-difference (DiD).[207] Instead of relying on only a few comparison areas, it uses a more complex model. As a result, it may be less susceptible to *bias* than a DiD design.[208] Synthetic control can also be used within a DiD design: if a suitable comparator was not available, a synthetic one can be created as a weighted average of a number of comparison *units*.

This design can be used to investigate impacts at the aggregate level too: across regions, states or countries. There have now been many different kinds of policy innovations tested, such as the impact of Japan's nuclear power facilities on citizens' income levels.[209]

Synthetic control relies on an understanding of which factors influence the outcome we are interested in, and reliable data on those factors across the whole 'donor pool', as well as for those *units* (in our case California) who did receive the innovation. If we do not have this knowledge, then the synthetic control may not be as comparable to the intervention group as we would like; as a result, the difference between them is not because of the innovation. One way to check this is to study whether the synthetic control and intervention groups appear to agree on pre-intervention outcomes. In the study on tobacco in **Box 16**, the authors began their analysis with data from 1970. This allowed them to build a model that was sensitive to factors that might influence change over a long time period, and use that to construct the synthetic control. The deviation of 'synthetic' California away from 'real' California can then be more readily separated out – as **Figure 20** shows.

Figure 20: Synthetic California cigarette sales per capita



Source: From The Urban Institute, 2017

The authors of the tobacco study argue that a key strength of this method is that it is transparent and data-driven. Anyone can access the data in the free software Synth, and other analysts can use it to run tests and interrogate results to see how well they stand up to scrutiny. In 2016, another team created a complementary software package that aims to build on Synth that's also available online, called synth_runner.[210] While its proponents value its transparency, the flip side of its use of machine learning is that can be less transparent – and harder to understand – for non-experts and decision-makers.

Synthetic control is an emerging approach: we have lots to learn about how best to analyse results, and in what situations the results are reliable. The normal statistical tests that scientists use to interpret results – tests of 'statistical significance' – can't be used with Synth due to the artificial nature of the comparison. Statistics is in some sense the study of natural variation, and it can tell you when a variation is so big that it is 'un-natural'. Any comparison involving a synthetic control is already un-natural, so the usual statistical rules do not apply. Analysts are still working out the best way to robustly test the findings of synthetic control.[211] We don't advise policy evaluators to use synthetic control without a thorough understanding of its limitations. The statistics behind synthetic control is complex; if you are considering using it, do seek expert statistical advice.

USP: Pre-experiments investigate what changes when an innovation is introduced, using only a single group. Useful for exploratory aims, formative evaluation (improving what you do) and trying out new ideas, rather than robust impact evaluation.

Pre-experiments 'perturb and observe': they make a change in the world, and seek to understand what effect it has had, using a single group. Like the other designs in this inventory, they investigate the *counterfactual*: 'what would have happened to those people who received this innovation, if the innovation had not taken place?'. Instead of using a control or comparison group, pre-experiments use individuals' past outcomes to estimate the *counterfactual*. They compare people in the present, who have received an innovation, with those same people in the past, before they received it. This approach is less effective at assessing causality, so pre-experiments are not usually helpful for establishing whether or not our idea is effective. But, they are useful designs when we want to probe and discover, shape new hypotheses, and flesh out novel ideas by trying them out in practice.

Some pre-experiments focus on two measurement points: before (pre) and after (post) an innovation is introduced. Others take repeated measurements across a time period, to get a more nuanced reading of change or as continuous innovation takes place. Some approaches, like rapid cycle testing, can be used for formative evaluation: improving what you do, and how you do it. Others, like prototyping, are focused on building a solution with a better chance of success through stakeholder engagement. Each design in this section faces threat from different kinds of *bias* (some of which are summarised in **Table 2**).

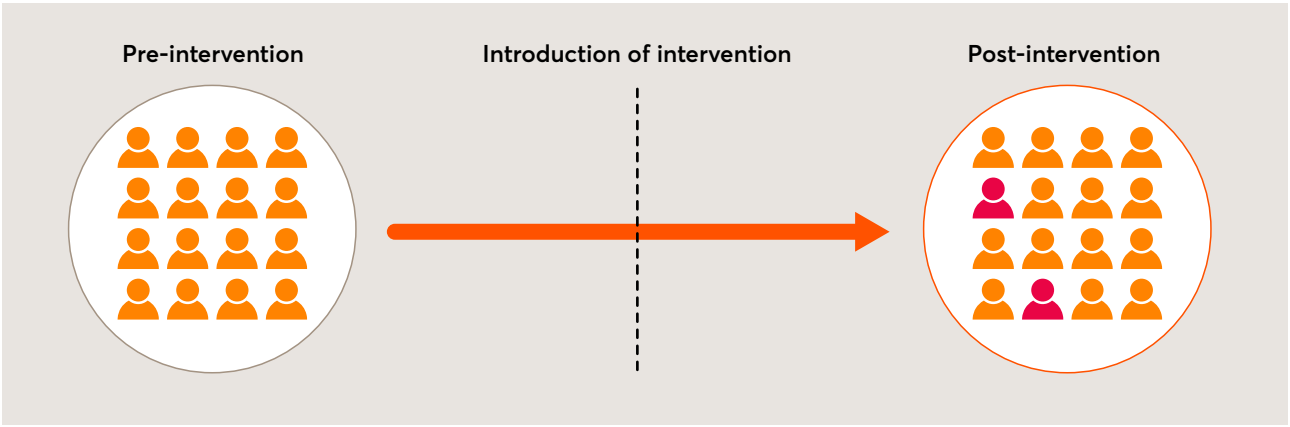
When we are aware of their limitations, pre-experimental designs can help us learn about the ingredients of a successful solution. They emphasise the groundwork that must be done to develop policy ideas with promise. Innovations developed with pre-experiments can be more rigorously evaluated with quasi-experimental or randomised designs.

3.16 Pre-post test

USP: The simplest design, compares outcome measures before and after an intervention.

The pre-post design is the most basic of all the approaches in this inventory and serves as a foundation for the pre-experimental designs in this section. In its simplest terms, a pre-post design takes measurements about outcomes for a group receiving an intervention before (pre) and after (post) the intervention takes place. It then calculates the difference between them, to see what's changed. A pre-post design can also be called a before and after test: it gathers pertinent information about participants and the outcomes that are relevant to the innovation before and after it's introduced. To take an example, if we wanted to find out whether a training programme helped adults develop new digital skills, we could administer an online test before they received the training and then after, to see whether they performed better after being trained. As with all kinds of trial, we'd want to know some basic things about participants – like their age, job title and previous experience perhaps; and we would aim to design a questionnaire that tested the outcome of a specified set of digital skills – like their competency with different kinds of software.

Figure 21: A pre-post test



There are some serious limitations to a pre-post design, which stem from the fact that it doesn't use a control or comparison group, and that (in its simplest form) it only takes measurements at two points. The lack of a control or comparison group means that it is very possible that any differences between outcomes could be due to a factor other than the introduction of an intervention. Changes in outcomes measurements could be due to different types of *bias* – like history *bias*, where an event or experience that isn't the innovation impacts on the group. Another is testing *bias*, which means that if a test (our online digital skills test, for example) is administered to the group before and after, they

might do better the second time around just because they've seen a similar test before, and know what you're looking for (but not because their general digital skills have improved). A pre-post design is generally more appropriate if:

- You do not expect much 'outside influence' on the outcome
- The duration of the trial is reasonably short
- Data about the outcome can be gathered fairly unobtrusively without 'training' individuals in the experiments (so perhaps using a data collection process that exists already, to avoid testing *bias*).

The main sources of *bias*, and the reasons why results from a pre-post test might be misleading, are detailed in **Table 2**.

Table 2: Threats to *internal validity*[212]

History	An event or experience impacts on individuals during the intervention period.
Maturation	Normal changes over time (like getting tired, older, or more experienced) affect the group in question during the intervention and affect outcomes.
Testing	The test administered to the group to gather pre-intervention measurements conveys new knowledge to them – and influences how they score on the second, post-intervention test.
Instrumentation/ reporting	Changes to the process of measurement between the pre and post tests (for example, inconsistencies in the data collection method used, who uses it, and how).
Regression-to-the-mean	If the sample was selected because they were extreme in some way (for example, if we selected roads to have additional road safety measures introduced on the basis that they were accident hot-spots) then we can expect their future outcomes to be less extreme, even if the treatment is completely ineffective. This *bias* is fatal to a pre-post design: if you have chosen who gets the intervention based on their past performance then a pre-post design will give you no indication at all whether the treatment worked and you need a comparison group.
Drop out	Individuals drop out of the intervention, creating a difference between pre-intervention and post-intervention measures that is not due to the innovation being evaluated.

For these reasons, a pre-post design is not a good option for understanding effectiveness and finding out what works. It could be a more viable option for developing an idea and refining it before it's more formally trialled. Before using a pre-post design, we should consider one-by-one the threats to the *internal validity* of the test. These are possible alternative explanations for observed results other than the innovation, and you should try to mitigate them. We only recommend this design if no other alternative is possible, as a last resort. It's worth considering too if it might be combined with other types of evaluation to generate more useful results. Depending on your aims, the other pre-experimental designs we explore in this section may be a better fit for developing your hypotheses, consulting with others on what a solution should look like, or trying out a new idea.

3.17 Rapid cycle testing

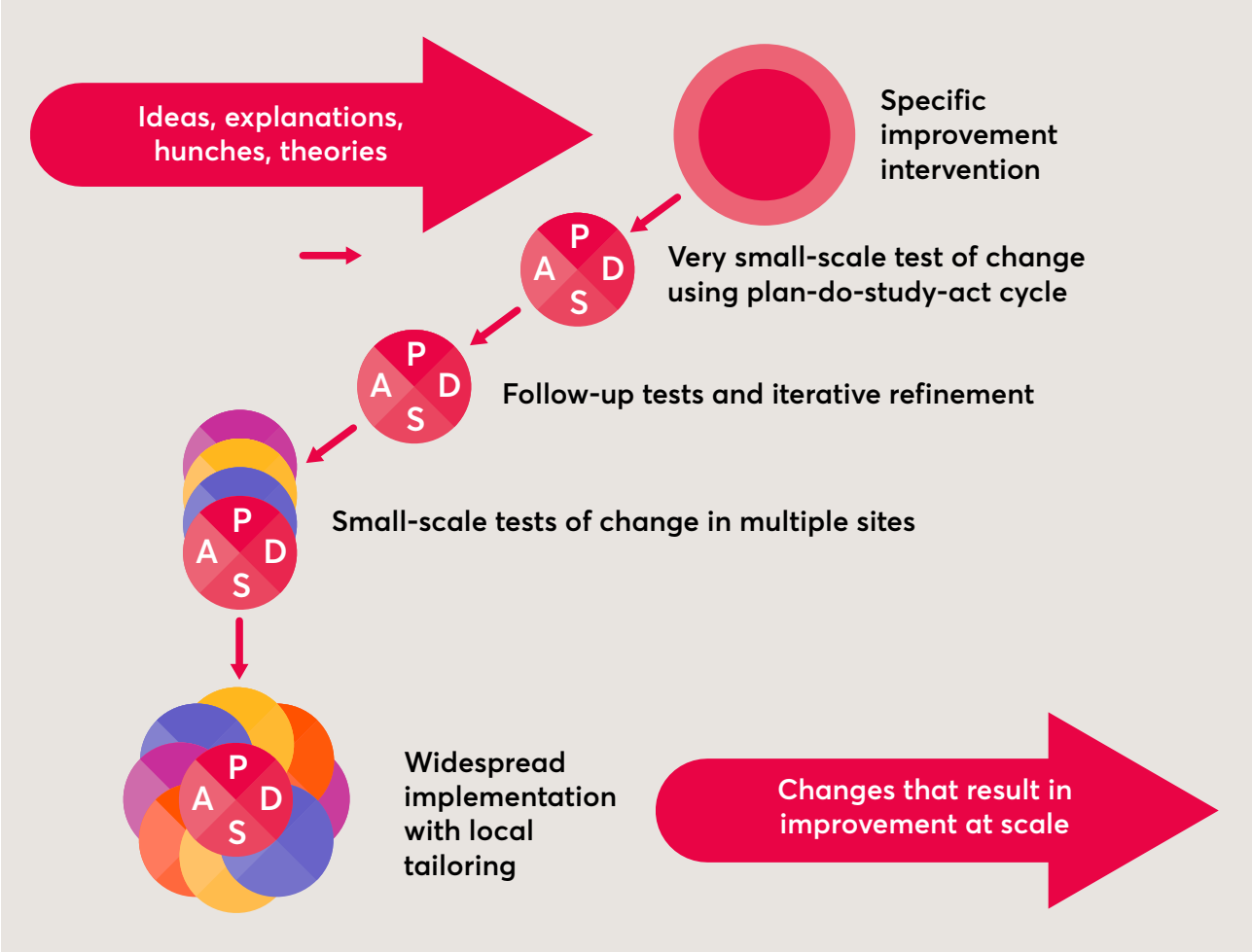
USP: Iterative experiments useful for local problem-solving, that create a rapid feedback loop between testing, re-design and re-testing.

Rapid cycle tests are a family of experiments being pioneered by early adopters in the US, Sweden, Canada and the UK. They draw on diverse sources of inspiration – like improvement studies and implementation science in health, and 'lean startup' and 'lean impact' approaches in the emerging social research and development (social R&D) sector.[213] Rapid cycle tests take a more flexible approach to evaluation, building on basic pre-post designs to create an iterative approach. They take multiple measurements over a longer time period, as innovations are re-designed and re-tested, and focus on solving the problems that face decision-makers locally.

Rapid cycle tests are conducted in lots of ways, but most are underpinned by 'plan-do-study-act' (PDSA), a template for involving practitioners, leaders and researchers in developing and testing innovations that was developed in healthcare improvement. Rapid cycle tests build on this practitioner-led model to create series of small experiments. **Figure 22** shows this iterative approach, in which multiple small trials are conducted, and results are built upon over time. Other ways of doing rapid testing are usually a twist on this design: Dartington Service Design Lab, which is leading the way on doing rapid testing in new policy areas in the UK, uses a five-step process that aims to combine new evidence with user-centred design.[214] Rapid tests emphasise continuous learning and often require the whole organisation to plan, collect and analyse data. Some approaches rely on frontline staff testing things out and have little or no input from expert evaluators. Others draw on the expertise of both practitioners and researchers, and use more advanced measurement techniques.

Rapid cycle testing obviously lacks the rigour that more formal experimentation methods gain by using a comparison or control group, but this may not matter very much depending on the context. If you are running an online platform for example, you control the entire user journey and, if it is not particularly susceptible to outside events, then any change you observe post-treatment can reasonably be attributed to the change that you made. This is particularly true given the rapid nature of the experiments: there has been little time for anything in the outside world to change.

Figure 22: Using rapid cycle tests for improvement



Source: From Greenhalgh and Papoutsi, 2019

Some of the most developed examples of rapid testing can be found in the US. At Harvard University, the Centre on the Developing Child is testing out ideas highlighted by America's National Scientific Council on the Developing Child, a multi-disciplinary collaboration that works to promote the science of early childhood development.[215] Through a project called Frontiers of Innovation, researchers and practitioners work together to turn ideas from research into interventions for children and families. One example of this is Learning Through Play, a programme developed with researchers at the University of California that draws on science about how children develop 'executive function' – the cognitive skills that help them pursue tasks and goals – in their early years.[216] The researchers created a programme that supports adults and children to play games together to develop these skills. Rapid cycle testing will often start with a small feasibility test with about ten families, then be iterated

with 20 or 30, and then, if an innovation shows promise, tested with a bigger sample and a control or comparison group. In this context, 'promise' does not usually mean a statistically significant improvement (the experiment is too small for that): any improvement after the intervention would be considered evidence of promise, and grounds for further testing.

Other organisations are using rapid tests to innovate in public services. Nesta's **People Powered Results** team runs 100 Day Challenges that search for new ways improve health and care services by bringing together multi-disciplinary teams.[217] Like others, they look at how changes to the practice of service delivery can create a more effective system. The US Centre for Employment Opportunities, which supports people who have been in prison, is using rapid tests to explore having staff with experience of the criminal justice system deliver their support programme. Based on evidence that suggests that people with lived experience might be better at building relationships, they are providing special training to these staff mentors, and running rapid cycle tests to see if the innovation can improve how engaged participants are, or reduce drop-outs from the programme.[218]

Many applications of rapid testing (including the one above) focus on adapting projects to make them more appropriate for a specific group of people or in a particular context.[xi] The Family Nurse Partnership (**Box 17**), a home-visiting programme providing support for first-time young mothers, is exploring how this might be done at a national scale in the UK, adapting a big evidence-based programme originally created in the US.[219]

Box 17: Adapting the Family Nurse Partnership

The Family Nurse Partnership (FNP), founded by David Olds, was the first evidence-based programme for families taken to scale in England. The programme is committed to robust evaluation to test whether good outcomes can be achieved in different contexts outside of the US. The programme, which is delivered under licence, provides support for first-time young mothers from trained nurses or midwives who visit between pregnancy and the child's second birthday.[220]

In 2016, the programme was evaluated with a randomised controlled trial, and found to have little added benefit over normal care.[221] Because the programme has an internationally recognised evidence base – including three RCTs in the US – the FNP National Unit decided to customise the programme to make it more flexible, personalised and cost-effective in the UK. They partnered with Dartington Service Design Lab on an ambitious new project called FNP ADAPT – Accelerated Design and Programme Testing.[222]

FNP ADAPT draws on improvement approaches to adapt, test and learn about the FNP programme, while respecting its strong evidence base. Its aim is to identify changes that enable FNP to better meet the needs of families. At the outset, the project brought together a multidisciplinary team of FNP nurses, supervisors and commissioners from each participating area, alongside researchers, to co-produce changes to the programme. Participants worked together to develop logic models, as well as 'dark logic' models to map possible unintended harm that might result from the changes – for example, that new content might mean less time for other aspects of the intervention.[223]

Routine data collection in FNP – a licence requirement – meant that the team were able to use (and build on) a good existing quantitative dataset.[224] Both quantitative and qualitative data helped to inform decisions about whether innovations should be kept, tested further, or abandoned.

Rapid cycle tests require good quality data which is easily and quickly accessible to monitor the effect that changes have. Being able to test innovations this way requires having good 'baseline' measurements – that is, a good knowledge of the service, system and participants before changes are made, as well as during and after. They also require training for and commitment from staff across the organisation, to monitor and test innovations.[225] There may be certain kinds of social policy intervention that are more amenable to rapid testing; NSPCC, for example, is using rapid tests to evaluate place-based approaches, like its Together for Childhood initiative.[226] Rapid tests may be a particularly good fit for projects like these, which foreground local, community-centred knowledge creation.

At the moment, we don't have strong evidence to show that rapid cycle testing is consistently effective at improving the impact of innovations in the long-term. One challenge is that rapid testing can look very different across projects and sectors, so it's hard to work out what might make them effective in general.[xii] A second limitation is that, in some cases, we don't know what the unintended consequences of making changes to an existing programme might be. This is a question that researchers and practitioners in public health have been grappling with, and they are creating new frameworks, guidelines and tools to help decision-makers make safe and successful changes to existing projects and policies.[xiii]

There is potential to combine rapid cycle tests with 'nimble RCTs' or 'hybrid designs' – randomised experiments that we looked at in Sections **3.3** and **3.10** respectively. The use of these designs would give organisations more certainty about the effects of rapid innovations and the impacts they have for the people they work for, while still maintaining the pace of innovation. The result is similar in spirit to the A/B test (see Section **3.4**) which is standard practice in industries where innovation and optimisation is routine, such as web design.

3.18 Prototyping

USP: An approach to trying out ideas in practice before an innovation is offered to recipients, to get feedback and input from stakeholders, improve the idea and increase its chance of success.

Prototyping is focused on front-loading the risk of failure: by trying out ideas as early as possible, it aims to make small failures happen at the beginning stages of a project, so that the idea which is more formally trialled has a better chance of success. It can also be 'human-centred', focused on user and stakeholder engagement, a way to gather feedback, engage partners and communities, and create buy-in for a new idea. If an idea proves to be unworkable, this is found out before significant resources have been committed; if it needs to be improved then this is discovered at a point where a meaningful decision can be made as to whether to invest in improving it or whether it is better to try something else.

In social science terms, prototypes are not experiments because they don't involve structured testing and evaluation. Nevertheless, prototyping has an important role in helping organisations to think and act more experimentally. They are part of the puzzle of how to embed experimental mindsets, as well as methods. Nesta's former CEO Geoff Mulgan has argued that there is a craft to developing ideas experimentally: too much too soon can kill a good idea, while too little too late risks social irresponsibility.[227] Prototypes encourage us to reframe potential solutions as hypotheses and approach ideas with a testing mindset. Nesta's publications *Designing for Public Services* and *Prototyping Framework* aim to make prototyping practical and accessible for government practitioners.[228]

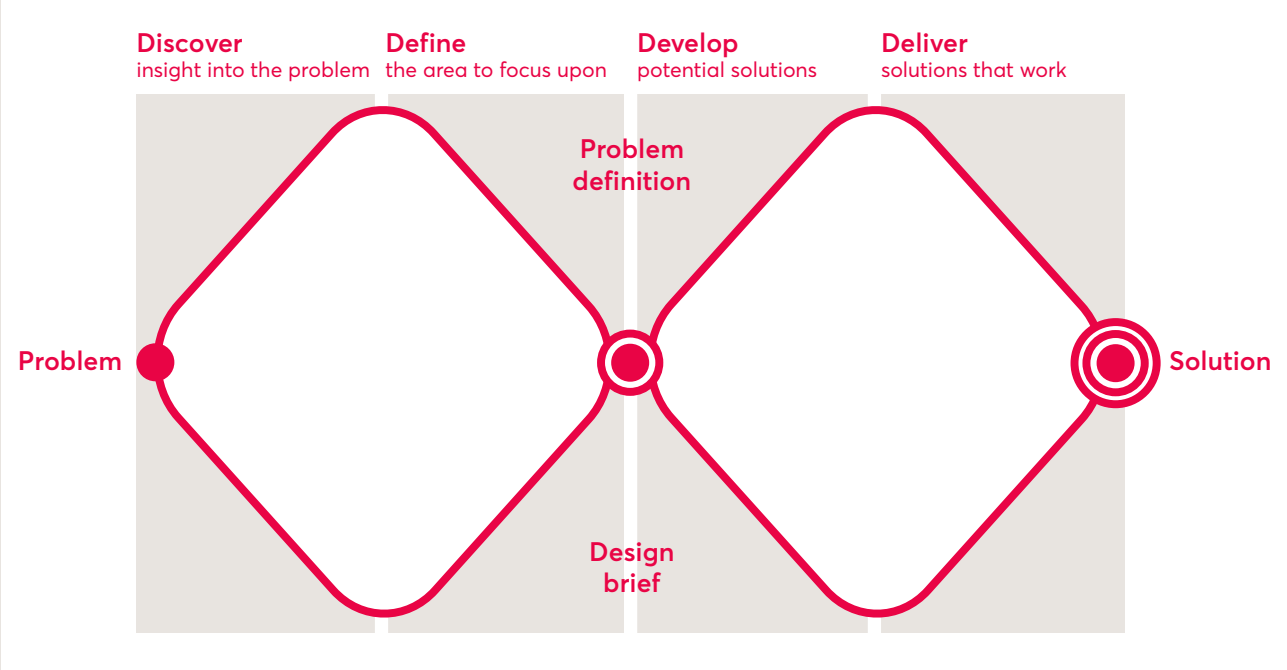
A prototype follows the 'double diamond': an iterative process that moves from a version with little detail or functionality (like a rough draft that illustrates the idea) to a version with much more detail and functionality (giving test users a better sense of how it is intended).[229] What's learnt through early iterations helps create a more refined solution that's eventually tried out in practice.[230]

Box 18: The design 'double diamond'

Divided into four distinct phases – discover, define, develop and deliver – the double diamond is a simple visual map of the design process.

It proposes that creative processes involve a number of possible ideas being created ('divergent thinking') before refining and narrowing down to the best idea ('convergent thinking'), and this can be represented by a diamond shape. But the double diamond indicates that this happens twice – once to confirm the problem definition and once to create the solution. This means that ideas are developed and refined a number of times, with weak ideas dropped in the process.

Figure 23: The Design Double Diamond



Source: Design Council, 2018

'Exploratory prototypes' investigate how a solution might look and feel through low-*fidelity*, low-resource models. A product can be created in cardboard form, a website through a hand-drawn wireframe, a service through role play.[231] These mock-ups can be shared with colleagues, stakeholders and users to elicit feedback.[232] The ideas for exploratory prototypes like these are usually based on pared-down ethnographic methods, where a designer or team observes or interviews service users, or alternatively a workshop or focus group.

'Live prototyping' covers later rounds of iteration, when a product or service is trialled at a small scale with users. Live prototypes aim to make operational errors and design flaws obvious and easier to avoid.[233] Unlike a full pilot, a prototype is often a simulation: the new service or website will not be fully up and running and may not be implemented in its normal setting (though it is preferable if it is). Online prototypes are often fairly straightforward to implement and evaluate; the UK Government Digital Service, for example, has a downloadable Prototype Kit which builds new versions of online government services.[234] Online prototypes can be evaluated with A/B tests, an online randomised trial that we covered in Section **3.4**. Other types of service innovations can be trickier to develop and evaluate. In Chile, Laboratorio de Gobierno's 'We are Community' project provides a good example of how an innovation lab used a quasi-experimental design (see Sections **3.12-3.15**) to robustly evaluate a series of prototyped innovations that aimed to help citizens in Santiago feel safer in their neighbourhoods.

Box 19: ¡Somos Comunidad! Making Santiago safer for citizens

Somos Comunidad (We are Community) aimed to find new solutions to citizens' perception of insecurity and risk in Santiago. Chile's Laboratorio de Gobierno (LabGob) took a robust approach to evaluating prototypes that were developed collaboratively with the public.

Research showed that Santiago's citizens had a much higher perception of the likelihood of crime happening than was apparent in the official statistics. To explore the problem, LabGob ran a consultation involving more than 300 people in workshops including police and local government, academics, community leaders and citizens. Several different solutions were proposed and went through iterative prototyping by teams supported by the government labs' practitioners. One solution was online: an app to report crime or fear of crime, developed with the police. Another was social: the creation of community 'Task Forces' who worked with local mayors. These Task Forces explored municipal solutions to anti-social behaviour and involved community police, security experts, and local community members, and tried different ideas out in six local areas. Each Task Force also trained citizens in how to use the online reporting app safely.

While each community was given autonomy to develop its own solutions, the programme was evaluated with the help of the Inter-American Development Bank, using a quasi-experimental method that compares communities to an untreated comparison group. At the time of writing, results were not yet in but Beatriz Hasbún, LabGob's Learning Experience Designer, is open about the decision to evaluate: *"This is a large-scale project and we wanted to work with an independent partner. There are so many different factors that affect people's perception of security – we felt it was crucial to use a comparison group."* If LabGob sees an impact on outcomes in the evaluation, it will support more neighbourhoods to test the programme. If not, solutions will be improved or discontinued.

Some ways of doing prototyping emphasise the 'quick and dirty' over a need for good research design. Prototypes can and should draw on existing research to understand the problem area they are trying to address, like exploring what's been done before, to see which ideas are most promising. Ad hoc or one-off field visits can only tell us so much because what's seen in one place, on one day, is unlikely to be representative. And while doing prototyping in workshops or user feedback sessions might provide a useful sounding board, it's often unclear how groups are managed or who gets invited. It's likely that different kinds of *bias* might be at work, like 'groupthink', which can lead to poor group decision-making.[235]

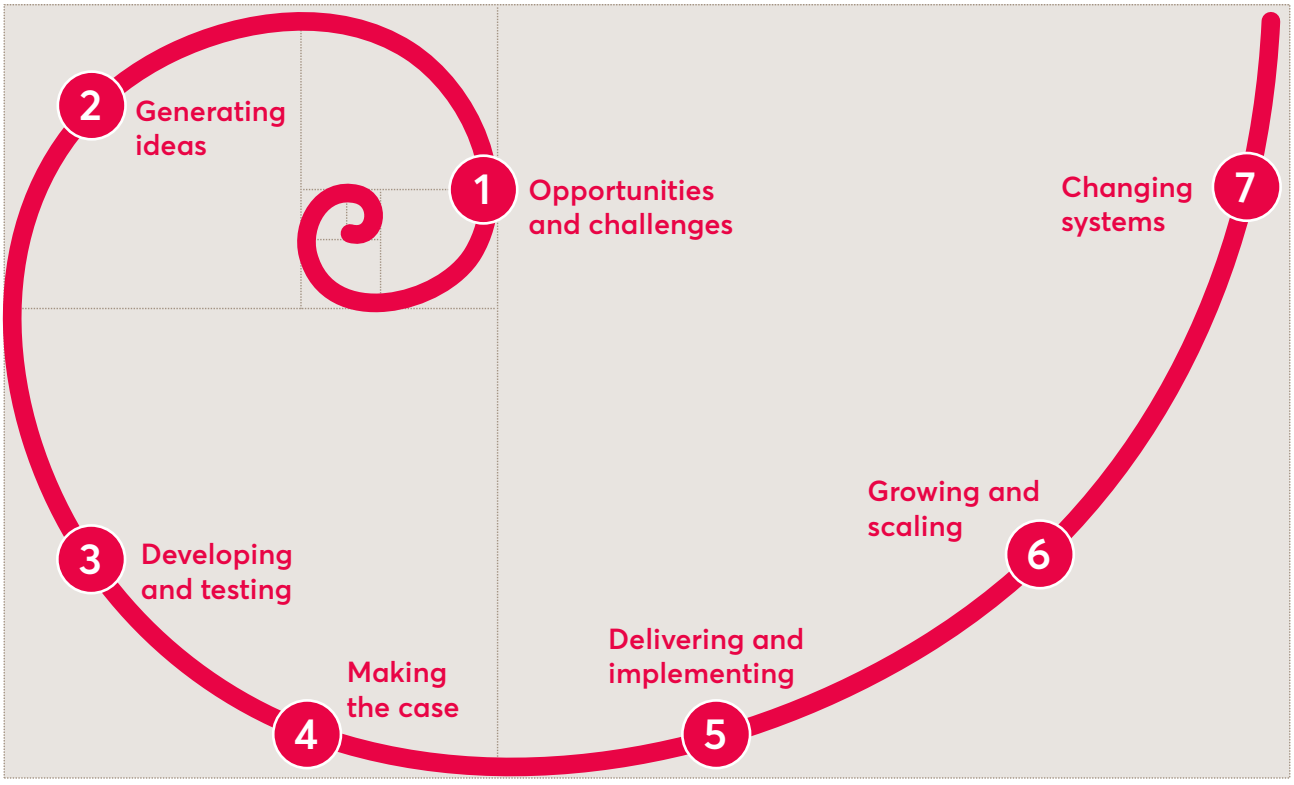
When it comes to live prototyping, there is still relatively little published literature on how these tests work in practice, and the process is likely to vary significantly from project to project. One question is whether prototypes are being evaluated in a way that would really allow them to fail, or whether they could fall into the trap of 'single loop learning', where preferred options are given the best chance of success while other ideas are not given a similar chance to show promise.[236] A possible solution to this could be to do more 'parallel prototyping' where multiple solutions are developed and tested by multiple designers, and to share learning among organisations about what great prototyping looks like.[237] US designer Bruce Hanington, co-author of the book *Universal Methods of Design*, points out that design should aspire to research excellence in order to bring 'relevance' together with 'rigour'.[238]

Conclusion

Not every experiment is right for your needs and strengths. For instance, running quasi-experimental designs (Section **3.12-3.15**) requires complex statistical know-how, whereas there are simple online experimental packages that anybody can use to run A/B tests on the digital delivery of services. Crucially, what's right for you will depend on what you want to learn about. The important principle is 'horses for courses': choosing the kind of experiment that is suitable for your challenge and environment. To help you here, **Table 1**, summarises all 18 experiments in our inventory.

When making your choice, it is worth thinking about how far you are on your journey of developing an idea. Some early ideas may just not be ready for full-blown RCTs or QEDs. Prototyping may be more what you are looking for (Section **3.18**) and emphasises learning with stakeholders about how an idea can be made a good fit for a community or group of users. If you want useful insight quickly, nimble RCTs offer pragmatic, operational learning and formative (improvement-focused) evaluation. Nesta runs **100 Day Challenges** in healthcare to focus minds, and focus results. When problem-solving requires pooling the expertise of teams more effectively, rapid cycle testing can combine a collaborative model with a degree of evaluative rigour (see Section **3.17**). These approaches can still be grounded in the best available science – and we found great examples of this work from organisations like the Dartington Service Design Lab in the UK and the Harvard Centre on the Developing Child in the US. Prototyping can also be combined with impact evaluation designs, as we saw in the example from Chile's Laboratorio de Gobierno (see **Box 19**).

Figure 24: Nesta's Innovation Spiral



Source: Nesta

However, if you have significant funding, a developed policy intervention, and are about to embark on a national roll-out, a more robust experiment like a randomised trial will give you the evidence you need to make crucial investment decisions. There are many designs out there, suited to a wide range of contexts, sectors and innovations (covered in Sections **3.1-3.11**). If a policy or programme is already underway, then you may have to evaluate things retrospectively (not prospectively, as with RCTs or pre-experiments). If this is the case, non-randomised and quasi-experimental designs (QEDs) will help – although it's worth keeping in mind that they rest on assumptions that may or may not be testable.

The issue may not, however, be time, or how far developed your innovation is. You might want to ask different sorts of questions of your innovation: instead of 'did this work', the question could be 'how and where will this work in practice?'. If that is the case, then hybrid trials (Section **3.10**) or realist trials (Section **3.9**) – drawing on qualitative research and openness to theory – are your best bet. Multi-site trials too (Section **3.8**) point to how we might grow the evidence base by testing portable theory. It's also worth keeping in mind different approaches to experimental design, particularly the stepped wedge and wait-list designs (Section **3.6**), which allow you the option of ensuring that all possible recipients can benefit in the end. When randomisation isn't possible, there are valuable alternatives: differences-in-differences, propensity score matching, and regression discontinuity (Sections **3.12-3.15**).

Whatever experiment you choose, we hope that you avoid the common myths and misunderstandings – like those about ethics, costs and speed – and do not use these as reasons not to experiment and improve. We have shown that there is a diversity of trials, and that some design innovations have helped to answer critics (see our history of experimentation in **Section 1** along with **Annex B**, which covers both criticisms and responses).

No approach is perfect, and no design offers easy answers. Many innovations fail to make a difference; making better policy takes hard work and investment. According to Jon Baron, Vice President of Evidence Based Policy at the Arnold Foundation, and a long-standing advocate of randomised experiments, most projects evaluated by an RCT show little impact. In education, 90 per cent of interventions evaluated in RCTs by the US Institute for Education Sciences had weak or no positive effects.[xiv]

This lack of progress should not, however, depress us. It is a reminder of how difficult it is to make a difference (and how unwise it is to assume that just because an innovation is new and sounds attractive, that it will improve outcomes in practice). It has been dubbed Rossi's Iron Law of evaluation, after the American sociologist Peter Rossi, who declared that: *"The better designed the impact assessment of a social program, the more likely is the resulting estimate of net impact to be zero."*[239] This law means that the more technically rigorous we are in how we evaluate, the more likely are its results to show no effect. As former Australian Government minister and advocate for RCTs, Andrew Leigh has argued:

> *"Rossi's Law doesn't mean we should give up hope in changing the world for the better. But we ought to be sceptical of anyone peddling panaceas. The belief that some social programs are flawed should lead to more rigorous evaluation and patient sifting through the evidence until we find a program that works."*[240]

To do this, we need to draw on many different tools and drive innovation in how we learn about making change. 'Sifting through the evidence' is not solitary work. Organisations and teams must take an open approach, that allows us to strengthen collective expertise, and identify important lessons more effectively. We must also work across sectors, to link up what we learn above and beyond our own areas of specialism – including with the wealth of knowledge housed in universities and research organisations.[241] There are many areas of policy where experiments are likely to be critical to progress. Fields like renewable energy and Green Deals may be ideal spaces for learning experimentally about how to improve efficiency in the energy sector, or what will work to change behaviour and help businesses take up low carbon practices. Similarly, devolution and regional policy changes, as well as future international collaborations, may provide more opportunities to learn about what works, when, and for whom.

As innovators seeking social good, we know how hard it can be to move the dial towards sustained success. Yet with humility and hard work, we have a duty put our ideas to the test – to stop doing what doesn't make a difference, find out what does, and understand how we can improve people's lives.

Section 4. Useful resources

There are now many resources available to support experimenters. If you are reading this in the UK, a dedicated Trials Advice Panel sits at the Cabinet Office to provide support and advice to government, and for others such as in NGOs, the Test+Build platform, run by the Behavioural Insights Team, provides support on online experimenting.

To encourage the adoption of technology and modern management practices in small and medium-sized enterprises (SMEs), the UK Department for Business, Energy and Industrial Strategy is trialling a new approach to funding experiments through its **Business Basics Fund**, supported by Innovate UK and Nesta's **Innovation Growth Lab**. We now have a network of 12 What Works Centres, many of whom curate and communicate the results of experiments for frontline professionals and decision-makers.[242] This trend is not unique to Britain.

Below is a list of some of our favourite, freely accessible resources:

Bit by Bit: Social Research in the Digital Age (2017) by Matthew Salganik. See chapter 4, Running Experiments. A digital copy is available free online at bitbybitbook.com

Designing for Public Services (2017) from Nesta and IDEO covers prototyping and other design-based approaches to innovating in services.

Developing and Evaluating Complex Interventions (2006) from Medical Research Council. An excellent source of guidance and advice on evaluating complex projects, with an update due for publication in 2019/2020.

Evaluation: What to consider (2015) from The Health Foundation covers rapid evaluation, along with approaches to evaluating improvement in health. More insight and blogs can be found on the webpages of its Improvement Analytics Unit.

Experimentation Toolkit (2018) from Innovation Growth Lab (IGL). This practical toolkit is designed to build an understanding of how adopting an experimental approach can be used to make policies more effective.

Getting to Moonshot: Inspiring R&D practices in Canada's social impact sector (2016) from SiG Canada covers promising examples of social R&D practice.

Goldilocks Deep Dive: Introduction to rapid-fire operational testing for social programs (2016) from Innovations for Poverty Action (IPA). A short guide on running Nimble trials.

Impact Evaluation in Practice (2011) from the World Bank. Clear and accessible textbook on doing randomised and quasi-experimental evaluation, with examples.

Impact evaluation methods: What are they and what assumptions must hold for each to be valid? (2015). A summary factsheet from J-PAL North America.

Mastering Metrics: The paths from cause to effect (2014) by Joshua Angrist & Jörn-Steffen Pischke. Excellent guidance on quasi-experimental designs.

Measuring Impact by Design: A guide to methods for impact measurement (2019) from Impact and Innovation Unit, Government of Canada. A new reference guide for those involved in the design, delivery, procurement or appraisal of impact measurement strategies in Canada and elsewhere.

Prototyping Framework (2013). The prototyping process outlined in this toolkit was developed by Nesta and thinkpublic.

GOV.UK Prototype Kit. This kit rapidly creates HTML prototypes of GOV.UK services.

Rapid Cycle Design and Testing: The what and the why (forthcoming, 2020) from Lowther, K., Simpson, D., Green, F., Morpeth, L,, Axford, N., and Hobbs, T. at the Dartington Service Design Lab. Some blogs and resources on rapid cycle testing are available on Dartington Service Design Lab's website.

Running Randomised Controlled Trials in Innovation, Entrepreneurship and Growth: An introductory guide (2016) from Innovation Growth Lab. IGL's step-by-step guide on running randomised trials.

Randomised controlled trials: Gold standard or fool's gold? The role of experimental methods in voluntary sector impact assessment (2016) from NCVO/Charities Evaluation Service.

Quality & Improvement: Theory & Practice in Healthcare (2008) from NHS Institute for Innovation and Improvement & Manchester Business School summarises some of the improvement approaches discussed. More resources can be found at NHS Improvement online.

Quasi-Experimental Designs and Methods (2014) from UNICEF. A brief for the charity sector on QEDs.

Test+Build and Predictiv platforms from the Behavioural Insights Team (BIT) allow you to build experiments online, supported by behavioural science.

Testing Social Policy Innovation (2014) from the European Commission. This guide covers both randomised and quasi-experimental designs and can be found on the Alliance for Useful Evidence website.

Test, Learn, Adapt: Developing public policy with randomised controlled trials (2014) from Haynes, Service, Goldacre & Torgerson. A classic guide on RCTs from the Behavioural Insights Team in collaboration with Ben Goldacre, author of Bad Science, and David Torgerson, Director of the University of York Trials Unit.

The Magenta Book: Guidance for Evaluation (2011) from HM Treasury. UK Government guidance on what to consider when designing an evaluation.

The London School of Hygiene and Tropical Medicine's Centre for Evaluation provides resources on methodologies for evaluation in public health.

Using Randomised Controlled Trials in Education (2017) from SAGE. A quality resource from education specialists at Queen's University Belfast.

Annex A: Experimental jargon buster

Bias: In simple terms, *bias* is when ideas or evidence about something are skewed or unfairly weighted in one direction or another. In experiments, *bias* is a general term for any phenomenon which causes us to mis-estimate the effect of an intervention in a systematic way (for an example, see *selection bias* below). *Bias* can be present in experiments if they are not well designed and will affect the strength of findings; it may cause researchers to over- or underestimate the effects of an intervention, for example. *Bias* is usually something to be avoided, but in some situations, it may be better to accept some *bias* if the alternative is to have a very imprecise estimate (known as a *bias*-variance trade-off).

Causal inference: When we say 'intervention X causes effect Y' we are making a causal statement (literally, a statement referring to what causes what). With a sufficiently robust experiment (such as an RCT or regression discontinuity), we can make such statements with confidence if the results of the experiment support them. With weaker or more exploratory designs, such as the pre-post design, we cannot say that any change is caused by the intervention, so those designs are not considered to be causal. Sometimes, if a trial would be causal but some unexpected event introduces a significant *bias* or other threat to the validity of the trial, the trial is said to have 'lost causality', meaning that you can no longer make a causal statement about the effect of the intervention.

Confounder: Trials generally try to 'isolate' the effect of the treatment as much as possible. A *confounder* is anything that could have an effect on the outcome which we would mistakenly attribute to the treatment. This usually depends on the design as well as the phenomenon. For example, if there were a positive time trend in the outcome and we were using a pre-post design, then we could mistakenly attribute the increase to the intervention and the time trend is a *confounder*. If we were using an RCT or a DiD design, then the time trend would not be a *confounder*, since it also applies to the comparison group (it 'cancels out' and we do not attribute it to the intervention).

Construct validity: This is the extent to which the theoretical concepts that form the basis of an experiment are accurate – and so whether or not an experiment is really testing the *hypothesis* it aims to. This is tricky because experiments sometimes test abstract constructs, with contested meaning. In a quasi-experimental evaluation a few years ago on the UK social programme Troubled Families, evaluators faced a problem when they realised that although the programme claimed to be for 'troubled families', it had actually been piloted with poorer families. These are not the same thing, so it became very hard to test the programme's theory – and also likely that the pilot targeted the wrong people.[243] (It's worth noting though that a new evaluation from the Ministry of Housing, Communities and Local Government has addressed some of these limitations.[244])

Contamination: Sometimes called spillover or spillover effect, *contamination* is when the treatment has an effect on '*units*' outside the intervention group. These effects can be either positive or negative. If the intervention is some form of advice, then individuals who receive it may tell their friends outside the intervention group, affecting their outcomes. Alternatively, if the intervention uses messaging along the lines of 'you have been specially selected for…' then participants outside the intervention group may think they haven't been selected and aren't eligible, which reduces their chances of doing whatever you wanted them to do. *Contamination* is a source of *bias* and is to be avoided if possible, unless doing so would make the experiment prohibitively large.

Controlled experiment: These are experiments that allocate some people to receive an intervention and others to a control group. In a randomised experiment, people are allocated to different groups by chance. In quasi-experimental designs, a comparison group will be created using statistical methods.

Counterfactual: The literal definition of *counterfactual* is 'something that didn't happen'. In the context of an experiment, the word is usually used to refer to 'what would have happened to the treatment group if we hadn't treated them'. This is never directly observed (that's why it's called a *counterfactual*), and the quality of an experiment can be judged by how well the control or comparison group approximates the *counterfactual*.

Experimentation vs innovation: Experimentation and innovation are sometimes used interchangeably, but they are not the same thing. Innovation is defined in the Cambridge English Dictionary as a new idea or method, or the use of new ideas and methods in the development of products, designs or ideas.[245] Social innovation applies this process to social problems. The European Commission describes social innovations as 'new ideas that meet social needs, create social relationships and form new collaborations. These innovations can be products, services or models addressing unmet needs more effectively.'[246] Experimentation, on the other hand, is an approach or set of tools that innovators might use.

External validity: Sometimes called 'transportability' or 'generalisability', this means the extent to which the results of an experiment can be applied more generally, in other places and times than those in the experiment.[247] This depends on several things, like the kind of question the experiment addressed, the sample size and composition (in particular how representative the sample was of the wider population), and the generalisability of the context in which the experiment took place, i.e. if the experiment took place in a very unusual setting, it's less likely the results will be relevant in other settings.[248] It also depends on how much we know about the innovation in question, and how it aims to create change: do we understand the mechanisms? Did we only learn what works, or also why? It's worth keeping in mind that theory is portable.

Fidelity: This word is commonly used to mean faithfulness. In an experimental context, it describes the extent to which an intervention is implemented as its designers intended.

Hypothesis: A statement that is to be tested. If posed as a question, it is a research question (for example, 'what is the effect of taking an aspirin each day on a patient's stroke risk?'). Research questions are used in all kinds of research and will guide the research design, methodology, data collection method, and analysis approach used. In statistical testing, the null *hypothesis* states that an intervention has no effect (for example, 'taking an aspirin each day has no effect on a patient's chance of having a stroke') and the alternative *hypothesis* states that the treatment does have an effect. The alternative *hypothesis* may or may not specify more about the nature of the effect. For example, 'an aspirin each day reduces the stroke risk'; 'an aspirin each day changes the stroke risk' or 'an aspirin each day reduces the stroke risk by 20 per cent' are all valid alternative hypotheses. An experiment will either provide enough evidence to reject the null *hypothesis* in favour of the alternative, or it will provide insufficient evidence (the null *hypothesis* can never be 'proven', similar to how a court of law will not find a defendant 'innocent', only 'not guilty').

Heterogeneity (of treatment effects): This is about how an intervention might affect people differently: treatment effect *heterogeneity* is the extent to which a treatment has differential causal effects on different people or *units* of the experiment. These are important because different *units* might respond differently to different programmes or solutions. For example, households with high levels of energy consumption might respond more to an experiment testing the effect of an electricity-saving intervention than households whose energy consumption is already low.[249]

Internal validity: *Internal validity* is about whether the experimental process was performed correctly, in a way that makes the test administered valid within the context of the experiment. Good *internal validity* allows us to say with confidence that X caused Y, because we have eliminated A, B and C – and any other factors that might be messing with our test – within the structure of our experiment. This is one of the strengths of lab experiments: they enable precise and clear tests by controlling the environment of the experiment. *Internal validity* is why it's important to get the implementation of experiments on the ground right.

Natural experiments: These are events which replicate experimental conditions, but which happened naturally (or if they are man-made, weren't intended as experiments when they were implemented). For example, when a new policy was rolled out across the country it may have been implemented on a rolling region-by-region basis, with the next region being chosen randomly each time, because it was logistically impossible to roll out to everywhere at the same time. The roll-out happened in this manner for purely logistical reasons, but because it happens to be identical to the stepped wedge experimental design, we can treat it as though it were an experiment and use the resulting data to evaluate the policy. *Natural experiments* generally arise when some external factor causes the innovation to be implemented in an apparently random way.

Selection bias: A phenomenon where trials can produce misleading results because of how the experimental groups were sampled or observed. For example, if we are measuring the effect of a voluntary programme that citizens can enrol themselves in, we might consider comparing the later outcomes of those who enrol themselves versus those who don't. This introduces *selection bias*, because people who proactively sign up for programmes tend to be more motivated and may be different in other ways, and this has knock-on effects on their final outcomes. Even if the intervention didn't do anything, you would probably conclude that it did because the intervention and comparison groups were different.

Unit (also experimental unit): This usually refers to whoever or whatever we measure outcomes for. Not to be confused with a cluster. If, for example, we have a randomised trial where each school receives the same intervention, but the outcome of interest is pupil performance, then the pupils are the *units* and the schools are the clusters. If the outcome of interest was a school's Ofsted rating then the school would be the *unit*, since the outcome relates to the whole school rather than an individual pupil. This usage of the word *unit* is not universal, and some sources will use the word *unit* to mean what would normally be called a cluster.

(Un)observable: As the name suggests, a quantity or action is observable if it can be observed in a quantitative way (such as exam results, an individual's age, whether someone changes their energy provider and so on). Something is unobservable in quantitative research if it is an intrinsic trait, such as an individual's motivation or laziness. You may be able to make an imperfect measure of these traits with a survey, but there will always be aspects which are not captured with a numeric measurement. Unobservable quantities are a common source of *bias* in experimentation and the more robust experimental designs, such as the RCT, are able to reduce the negative consequences of unobserved variables.

Annex B: Common criticisms of RCTs and responses

Randomisation is unethical	The ethics of randomised experiments have been hotly debated. One criticism is that it's unethical to experiment on humans and to deny people an intervention that could help them. A common question here is: 'Why give a treatment that can help one group but deny it to the control group?'. In medicine, a common ethical principle in trials is 'medical equipoise', which states that: 'Trials can only be justified if there is genuine uncertainty in the expert medical community about the preferred treatment. A physician must have an equal state of uncertainty – or 'equipoise' – between the available options.' If we do not know whether the intervention works or is even detrimental to outcomes, it is ethically acceptable to randomise. If we know it works (for example, through the use of previous randomised trials or other evidence, not just theory or 'common sense') we should not be randomising. Wherever there are sensitivities, ethical approval should be sought from an ethics committee in a university or governing organisation in the relevant area.
Trials are expensive and time-consuming	A common misperception is that trials can be very expensive and time-consuming. The origins of this critique come from clinical trials for drugs that are time-consuming and expensive due to the regulatory burden and standards that the trials must meet. But there is nothing inherently expensive with an RCT in a non-medical setting. Online experiments like A/B testing (see Section **3.4**), or nimble RCTs (see Section **3.3**) can be fast and cost-effective. There is nothing inherently expensive in a trial compared to other types of evaluation. The resource-intensity can come from the intervention (such as a big new welfare programme), or the evaluation itself (such as surveying thousands of people). But none of that burden is unique to trials. Implementing policies which have no effect is also expensive and time-consuming.
Participants will think randomisation is unethical	This is a genuine problem and can reduce the generalisability of results if those agreeing to be randomised do not form a representative sample. The set-up of a trial should incorporate discussions with those who will be running it regarding the principles behind randomisation. If participants understand why they are being randomised they are more likely to take part.

Limited generalisability	It can be hard to generalise from a single experiment to the wider world. The unique context of one time and place means that it can be difficult to learn policy lessons to roll out nationally. The best way to respond to this is to replicate any trial in multiple locations. Replication is an important part of experimental science and social science. It's also possible to test interventions in multiple places to grow the evidence base (see multi-site trials, Section **3.8**), and to use more mixed-methods research (see realist trials, Section **3.9**). Local context will always be a critical factor in learning from experiments. But it's also an issue for any form of evaluation or any roll-out of a policy. In a recent piece online, Eva Vivalt has pointed out, *"It's not that RCTs have uniquely bad external validity – it's just that generalizing from any study to guess implications for different programs elsewhere on the planet is hard work."*[251]
We cannot practice blinding in research trials involving social policy	In a drug trial, patients, doctors and researchers can be made unaware of what treatment patients are receiving (a double-blind placebo trial). In research in other areas – such as education, welfare and crime reduction – this is rarely possible since the intervention is clearly visible to all concerned. We are, however, often able to blind those involved in the measurement of outcomes. The lack of blinding is a problem for social policy research trials due to the propensity for intervention group participants to 'try harder' (known as the Hawthorne Effect). However, other evaluation methods will rarely improve on this since they will also not operate blinding.
An RCT doesn't tell us how something worked – and for whom, and under what circumstances	In general, a trial does not tell us how the intervention works. True. It tells us whether an intervention has worked. This is a good argument for qualitative work happening in parallel with any trial that also seeks to clarify how the intervention is causing any measured effect. Some types of trials can also help us 'look under the hood' to see why something has worked, such as realist trials (see Section **3.9**), or 'how' something works using hybrid trials (Section **3.10**), as well as multi-site trials, that tell us more about 'who' (Section **3.8**). In some areas, like economics, trials may be more deeply rooted in theory, and explore the mechanisms driving results.
The complex interplay of systems and structures within social policy means that large-scale policy changes are difficult to trial with an RCT	The introduction of a new social policy or programme may involve a whole raft of different interventions, laws, regulations, funding, and changes to practice. Trialling the whole policy may be unfeasible. However, it could be possible to test some specific aspects of the policy, such as a change to welfare programme, or a new tax incentive. If that is not possible, it may just be that an RCT is unrealistic, and other types of evaluation are more practical, such as quasi-experimental designs (see Sections **3.12–3.15**), or pre-experiments (see Sections **3.16-3.17**) on rapid iterations in the roll-out of the policy. The Medical Research Council is leading the way on providing guidance on evaluating and experimenting in complex systems (see our list of resources in Section **4**). It is also worth noting that experiments are possible in challenging multifaceted areas, such as on 'complex contagion' and how innovations spread like social movements or new social norms (see **Box 8**, which covers the work of Damon Centola).

Source: Adapted from NfER, 2010[250]

Endnotes

Roman numerals

i. Definitions of what constitutes an experiment tend not be consistent. We follow William Shadish, Thomas Cook and Donald Campbell (2002); Martin Ravallion (2018) and Matthew Salganik (2017) who define an experiment as a kind of trial, with different designs and forms of control. Try William Shadish, Thomas Cook, and Donald Campbell, Experimental and quasi-experimental designs for generalized causal inference (Wadsworth Cengage Learning, Boston, MA, 2002).

ii. We discuss several of these debates in the section on 'The Basic RCT' in this report. For reference, see critiques from Lant Pritchett (2018) 'The Debate about RCTs in Development is over. We won. They lost.' Development Research Institute Blog, NYU. Available at: http://www.nyudri.org/events-index/2018/2/22/lant-pritchett-talk-the-debate-about-rcts-in-development-is-over-we-won-they-lost. Also, this article summarising some debates resulting from the Nobel Prize in Economics: Piper (2019) 'The Nobel went to economists who changed how we help the poor. But some critics oppose their big idea.' Vox. Available at: https://www.vox.com/future-perfect/2019/12/11/20938915/nobel-prize-economics-banerjee-duflo-kremer-rcts

iii. Advocates of RCTs are often termed 'randomistas'. In some cases, this is considered an insult – but many also use it positively, such as Andrew Leigh in his 2018 book, Randomistas.

iv. Many early attempts at experimenting in the UK failed because decision-makers jumped the gun on making policy before results were in. Gough and Breckon provide a review of early UK policy experiments in Breckon & Gough (2019) 'Using Evidence in the UK' in What Works Now, edited by Boaz, Davies, Fraser & Nutley. The Magenta Book is included in our list of further resources in Section 4.

v. Find out more about how NSPCC are learning about their services and explore their resources at NSPCC Learning: learning.nspcc.org.uk

vi. Our Using Research Evidence Practice Guide describes a theory of change. 'In the early stages of any intervention, it's important to logically describe what you do and why it matters coherently, clearly and convincingly. This is often referred to as a theory of change, and aims to give a: *"…clear, concise and convincing explanation of what you do, what impact you aim to have, and how you believe you will have it. It is a vital foundation of any programme, and a prerequisite for effective evaluation."* A theory of change is a useful way to be more explicit about what evidence you are using from others – and to be clearer about how you are going to get results.' Jonathan Breckon, Using Research Evidence: A Practice Guide (Alliance for Useful Evidence, Nesta, London, 2016) p16.

vii. One example of a natural experiment is a study of the 2008 Beijing Olympics. It looked at the impact of air pollution in the city, and how this affected health. In preparation for the Olympics, the Chinese Government introduced measures to improve the city's poor air quality. These measures were reversed after the Olympic period, creating a small window of better quality air in the city. A researcher used this window to compare the weight of babies born in this period to those born in 2007 and 2009, and found they were 23 grams heavier on average, suggesting that air pollution can interfere with how babies develop in the womb. Seizing the opportunity to run this natural experiment made it possible to study something that it would be unethical to study in a randomised experiment. This natural experiment generated important evidence about the harmful effects of pollution. Rich et al (2015) 'Differences in Birth Weight Associated with the 2008 Beijing Olympics Air Pollution Reduction: Results from a Natural Experiment' in Environ Health Perspect. 2015 Sep;123(9):880-7. doi: 10.1289/ehp.1408795.

viii. Taking the same example, if we decide to compare people over 65 who received a flu jab to individuals over 65 who didn't, this is unlikely to resolve all our problems. On the one hand, people who received a flu jab might be different to those who didn't in some way we can observe – maybe they were the targets of an advertising campaign or received a letter from their doctor. On the other, there could be differences that we can't necessarily observe or measure – perhaps those who didn't get a jab are less engaged with health services and avoid going to check-ups. This is likely to impact on their health long-term.

ix. The Campbell Collaboration is a central part of the evidence landscape worldwide. It has produced systematic reviews on social interventions since 2000 and has centres in Denmark, the UK and India. Find out more: https://campbellcollaboration.org

x. Traditional matching can mean running into problems creating matches and is problematic when using large amounts of data. For example: what if few good comparison individuals exist? Or if no matches are available? In these situations it's difficult to determine what constitutes a good-enough match.

xi. The UK's Health Foundation has for example pointed out programmes can't be seen as packages that can just be 'copied and pasted' to new locations; they need to be implemented in a way that's sensitive to context. See Dixon (2016) 'Spreading improvement: how to accelerate and the importance of archetypes' in Khan, The Future of People Powered Health, Nesta, London.

xii. One systematic review of the 'plan-do-study-act' method, for example, conducted in 2014, reported that many interventions using PDSA didn't actually comply with many of the method's core features, and were often poorly evaluated. Only 15 per cent of the studies in the review reported using quantitative data at monthly or more frequent data intervals to inform cycles of change and learning. Another, published in 2019, pointed out that, 'widespread challenges with low adherence to key methodological features in the individual projects pose a challenge for the legitimacy of PDSA-based QI'. Other ways of doing rapid cycle testing may face similar challenges. See the following reviews: Knudsen, S.V., Laursen, H.V.B., Johnsen, S.P. et al. Can quality improvement improve the quality of care? A systematic review of reported effects and methodological rigor in plan-do-study-act projects. BMC Health Serv Res 19, 683 (2019) doi:10.1186/s12913-019-4482-6 (Open Access); Taylor MJ, McNicholas C, Nicolay C, et al. (2014) Systematic review of the application of the plan-do-study-act method to improve quality in healthcare, BMJ Quality & Safety 23:290–298.

xiii. Many public health programmes are adapted by local teams. One review found that 62 per cent of US evidence-based programmes focusing on HIV/AIDS, mental health, substance abuse, and chronic illnesses had been adapted locally; see Cam Escoffery et al, 'A systematic review of adaptations of evidence-based public health interventions globally', Implementation Science 13, no. 125 (September 2018). How these adaptations are made is important: there's a risk that making changes to programmes could contradict their original evidence-base, making them less effective, or even causing harmful consequences. But, it's often not clear what the 'core elements' or 'active ingredients' of a solution are. In public health there are now frameworks to help organisations make safe adaptations, such as ADAPT-ITT, which provides guidance on adapting and testing evidence-based HIV interventions. Some projects, like US sex education programme Get Real, have specific guidance on 'green light', 'yellow light', and 'red light' adaptations. See Gina Wingood and Ralph DiClemente, 'The ADAPT-ITT model: A novel method of adapting evidence-based HIV Interventions', Journal of Acquired Immune Deficiency Syndromes 47 (March 2008) Suppl 1:S40-6, and the Get Real website: https://www.etr.org/ebi/programs/get-real

xiv. See Coalition for Evidence Based Policy (2013) Randomized Controlled Trials Commissioned by the Institute of Education Sciences Since 2002: How Many Found Positive Versus Weak or No Effects, Coalition for Evidence Based Policy. For employment and training programmes 75 per cent of RCTs commissioned by the US Department of Labour show weak or no positive effects. The same is true in business. According to the author Jim Manzi, 80 to 90 per cent of RCTs on new products and strategies run by Google and Microsoft have found no significant effects (Jim Manzi, Uncontrolled: The Surprising Payoff of Trial-and-Error for Business, Politics, and Society, Perseus Books Group, New York, 2012, pp. 128 and 142).

Numbers

1. Find out about our EdTech Innovation Testbed at: https://www.nesta.org.uk/project/edtech-innovation-testbed; and read about the Business Basics Programme: https://www.nesta.org.uk/blog/why-you-should-know-about-business-basics-programme
2. Jonathan Breckon, Better Public Services Through Experimental Government (Alliance for Useful Evidence, Nesta, London, 2015).
3. The Franklin Institute, Science and Learning Resources, 'Edison's Lightbulb' https://www.fi.edu/history-resources/edisons-lightbulb
4. Nicholas Maxwell, Karl Popper, Science and Enlightenment (London: UCL Press, 2017).
5. Attributed to Kurt Lewin in Charles W. Tolman (1996) Problems of Theoretical Psychology - ISTP p.31 and Johnson & Christensen (2016) Educational Research: Quantitative, Qualitative, and Mixed Approaches, SAGE Publications.
6. Jim Manzi, head of Applied Predictive Technologies claims in his book Uncontrolled that his firm is running RCTs for '30 to 40 per cent of the largest retailers, hotel chains, restaurant chains, and retail banks in America'. Jim Manzi, Uncontrolled: The Surprising Payoff of Trial-and-Error for Business, Politics, and Society (New York: Basic Books, 2012) p147.
7. Brian Christian, 'The A/B Test: Inside the Technology That's Changing the Rules of Business', Wired, 25 April 2012 https://www.wired.com/2012/04/ff-abtesting
8. Andrew Leigh, Randomistas: How Radical Researchers Changed Our World (Yale University Press, 2018).
9. Gerry Stoker, 'Exploring the Promise of Experimentation in Political Science: Micro-Foundational Insights and Policy Relevance', Political Studies 58, no. 2 (March 2010): 300–319. https://doi.org/10.1111/j.1467-9248.2009.00812.x p316
10. Alvin Roth, 'Introduction to Experimental Economics' in The Handbook of Experimental Economics, edited by John Kagel and Alvin E. Roth, 3–109. Princeton University Press (1995).

11. Amy Edmondson, 'Strategies for Learning from Failure', Harvard Business Review, 2011 https://hbr.org/2011/04/strategies-for-learning-from-failure
12. Kincaid, H (2012) Introduction: Doing Social Science, The Oxford Handbook of Philosophy of Social Science Oxford University Press.
13. Gerry Stoker, 'Exploring the Promise of Experimentation in Political Science: Micro-Foundational Insights and Policy Relevance', Political Studies 58, no.2 (March 2010): 300–319. https://doi.org/10.1111/j.1467-9248.2009.00812.x p316
14. Jonathan Breckon, Better Public Services Through Experimental Government (Nesta, London, 2015) p7.
15. Jen Gold, 'How Experiments Craft Better Policy: Inside the world's What Works teams', Apolitical, 2019 https://apolitical.co/solution_article/inside-worlds-what-works-teams
16. Conversations with Nicholas Chesterley and Myra Latendresse Drapeau, Government of Canada.
17. Health Canada's Experimentation Works (EW) Team, 'Health Canada's PRODigy experiment: An experience in learning-by-doing', Medium, 2018 https://medium.com/@exp_works/health-canadas-prodigy-experiment-an-experience-in-learning-by-doing-87598ab222c3
18. Experimentation Works (EW) is on Medium: https://medium.com/@exp_works
19. See Department for Work and Pensions press release on Universal Credit, 2013: https://www.gov.uk/government/news/universal-credit-progress
20. Whistleblowers have reported that the expensive IT system is riddled with design flaws and a January 2017 parliamentary inquiry into the system is ongoing. See Commons Select Committee, Work & Pensions, Benefits Sanctions, The role of Universal Support: https://publications.parliament.uk/pa/cm201719/cmselect/cmworpen/1667/166704.htm#_idTextAnchor002 Also see Patrick Butler, 'Universal credit IT system 'broken', whistleblowers say', The Guardian, 2018 https://www.theguardian.com/society/2018/jul/22/universal-credit-it-system-broken-service-centre-whistleblowers-say
21. Amy Edmondson, 'Strategies for Learning from Failure', Harvard Business Review, 2011. https://hbr.org/2011/04/strategies-for-learning-from-failure
22. EEF Blog, 'Magic Breakfast: A case study in scaling evidence for impact', Education Endowment Foundation, 2018 https://educationendowmentfoundation.org.uk/news/eef-blog-magic-breakfast-a-case-study-in-scaling-evidence-for-impact
23. Jonathan Breckon, Using Research Evidence: A Practice Guide (Alliance for Useful Evidence, Nesta, London, 2016).
24. Jonathan Breckon, Better Public Services Through Experimental Government (Alliance for Useful Evidence, Nesta, London, 2015).
25. Anthony King and Ivor Crewe, The Blunders of Our Governments (London: Oneworld Publications, 2013).
26. A. J. Macdonald, Guide to the Classical and other Long-term experiments, Datasets and Sample Archive (Rothamsted Research, Herts, 2018) https://www.rothamsted.ac.uk/sites/default/files/Web_LTE%20Guidebook_2019%20Final2.pdf
27. Ronald Fisher, The Design of Experiments (New York: Hafner Publishing Compan, 1971 – originally 1935).
28. Sebastian Heilmann, 'From Local Experiments to National Policy: The Origins of China's Distinctive Policy Process', The China Journal No. 59 (January 2008): 1-30 Retrieved from: www.jstor.org/stable/20066378
29. Jonathan Breckon, Better Public Services Through Experimental Government (Alliance for Useful Evidence, Nesta, London, 2015).
30. Judith Gueron and Harold Rolston, Fighting for Reliable Evidence (New York: Russell Sage Foundation, 2013).
31. Robert Brook et al, The Health Insurance Experiment: A Classic RAND Study Speaks to the Current Health Care Reform Debate (RAND Research Brief, RAND Corporation, 2006) https://www.rand.org/pubs/research_briefs/RB9174.html
32. Jonathan Breckon, Better Public Services Through Experimental Government (Alliance for Useful Evidence, Nesta, London, 2015).
33. The report was titled When Will We Ever Learn?, the quote is provided in a history of evidence in Howard White (2019) The twenty-first century experimenting society: the four waves of the evidence revolution Palgrave Communications p3. https://www.nature.com/articles/s41599-019-0253-6.pdf
34. Jonathan Breckon, Better Public Services Through Experimental Government (Alliance for Useful Evidence, Nesta, London, 2015) p19.
35. Stephen Kidd, 'The demise of Mexico's Prospera programme: a tragedy foretold', Development Pathways Blog, 2019 https://www.developmentpathways.co.uk/blog/the-demise-of-mexicos-prospera-programme-a-tragedy-foretold/
36. What Works Network, The Rise of Experimental Government: Cross-Government Trial Advice Panel Update Report (What Works Network and Economic and Social Research Council, London, 2018).
37. For a 2019 update on the work of these teams: Jen Gold, 'How Experiments Craft Better Policy: Inside the world's What Works teams', Apolitical, 2019 https://apolitical.co/solution_article/inside-worlds-what-works-teams
38. Find out more about Experimental Finland at: https://kokeilevasuomi.fi/en/frontpage
39. The French Government website on the 'Garantie Jeune' can be found here: https://www.service-public.fr/particuliers/vosdroits/F32700

40. Caroline Allard and Ben Rickey, British What Works Centres: What lessons for evidence-based policy in France? (Agence nouvelle des solidarités actives, 2017) p 86. https://www.alliance4usefulevidence.org/assets/Ansa_A4UE_whatworks_final_Full-report-standard.pdf
41. Shatha Alhashmi and Giulio Quaggiotto, 'Promoting experimentation in government – learning from Canada's experience', Nesta blog, June 2017 https://www.nesta.org.uk/blog/promoting-experimentation-in-government-learning-from-canadas-experience
42. Experimental Finland website: https://kokeilevasuomi.fi/en/frontpage
43. Albert Bravo-Biosca and Lou-Davina Stouffs, 'What are we learning from policy experiments to increase innovation and entrepreneurship?', Nesta blog, June 2018 https://www.nesta.org.uk/blog/what-are-we-learning-policy-experiments-increase-innovation-and-entrepreneurship
44. James Phipps and Lou-Davina Stouffs, 'Coming soon... the Vouchers Story', Innovation Growth Lab blog, April 2018 https://www.innovationgrowthlab.org/blog/coming-soon-vouchers-story
45. Dan Hodges, 'How experiments are improving the way we work', Innovation Growth Lab blog, August 2018 https://www.innovationgrowthlab.org/blog/how-experiments-are-improving-way-we-work
46. Find out more about Giving Evidence at: https://giving-evidence.com
47. Christian Bason, Design for Policy (Gower Publishing, 2014).
48. Geoff Mulgan, 'Mindsets and Methods: The 21st century curriculum for public servants', Nesta blog, April 2019 https://www.nesta.org.uk/blog/mindsets-and-methods
49. Find out more about States of Change at: https://states-of-change.org
50. William Trochim, 'Donald T. Campbell and Research Design', American Journal of Evaluation 19, no. 3 (1998): 407-409.
51. For an overview, see this report from the World Bank Group: Zeina Afif, William Wade Islan, Oscar Calvo-Gonzalez, Abigail Goodnow Dalton, Behavioral Science Around the World: Profiles of 10 Countries (English) (eMBeD brief, World Bank Group, Washington, D.C, 2019).
52. Available from the James Lind Library. MRC, 'Streptomycin treatment of pulmonary tuberculosis: A Medical Research Council investigation', BMJ 2 (1948): 769-78. https://www.jameslindlibrary.org/medical-research-council-1948b
53. See discussion in Harry Rutter et al, 'The need for a complex systems model of evidence for public health', The Lancet 390, no. 10112 (2017) 2602-2604.
54. Donald Berwick, 'The Science of Improvement', JAMA. 299 no.10 (March 2008): 1182–1184. https://doi.org/10.1001/jama.299.10.1182
55. Find out more about Edna McConnell Clark Foundation at: www.emcf.org
56. Vinod Rajasekaran, 'Getting to Moonshot: Inspiring R&D Practices in Canada's Social Impact Sector', SiG, 2016, p9.
57. Collins Dictionary Online: https://www.collinsdictionary.com/dictionary/english/experiment
58. James Druckman et al, 'The Growth and Development of Experimental Research in Political Science', American Political Science Review 100, no. 4 (2006).
59. Definitions of what constitutes an experiment tend not be consistent. We follow William Shadish, Thomas Cook and Donald Campbell (2002); Martin Ravallion (2018) and Matthew Salganik (2017) who define an experiment as a kind of trial, with different designs and forms of control. Try William Shadish, Thomas Cook, and Donald Campbell, Experimental and quasi-experimental designs for generalized causal inference (Wadsworth Cengage Learning, Boston, MA, 2002).
60. Jonathan Breckon, Better Public Services Through Experimental Government (Alliance for Useful Evidence, Nesta, London, 2015).
61. EEF, Classification of the security of findings from EEF evaluations (Education Endowment Foundation, July 2019) https://educationendowmentfoundation.org.uk/public/files/Evaluation/Carrying_out_a_Peer_Review/Classifying_the_security_of_EEF_findings_2019.pdf
62. Matthew Salganik, Bit by Bit: Social Research in the Digital Age (Princeton, NJ: Princeton University Press, 2017) p148.
63. Neil Salkind, 'Pre-Experimental Methods' defined in SAGE Encyclopaedia of Research Design, SAGE Publishing, 2010 https://methods.sagepub.com/Reference//encyc-of-research-design/n330.xml
64. Figure 3 is informed by: Neil Salkind, 'Pre-Experimental Methods' defined in SAGE Encyclopaedia of Research Design, SAGE Publishing, 2010 https://methods.sagepub.com/Reference//encyc-of-research-design/n330.xml; Sandesh Adhikari, '20 differences between Randomized Controlled Trial (RCT) and Quasi-experimental study design', Public Health Notes, 2018 https://www.publichealthnotes.com/20-differences-between-randomized-controlled-trial-rct-and-quasi-experimental-study-design; and Matthew Salganik, Bit by Bit: Social Research in the Digital Age (Princeton, NJ: Princeton University Press, 2017).
65. Elizabeth Lillie, Bradley Patay, Joel Diamant, Brian Issell, Eric Topol, Nicholas Schork, 'The n-of-1 clinical trial: The ultimate strategy for individualizing medicine?' Personalized medicine 8 no. 2 (2011): 161–173. https://doi.org/10.2217/pme.11.7
66. Taken from: Siobhan Campbell, Gemma Harper, 'Quality in policy impact evaluation: Understanding the effects of policy from other influences' (Magenta Guide: HM Treasury, DECC and DEFRA, 2012).

67. Paul Connolly, Ciara Keenan and Karolina Urbanska, 'The trials of evidence-based practice in education: A systematic review of randomised controlled trials in education research 1980-2016', Educational Research, 60 no.3 (2018): 276-291, DOI: 10.1080/00131881.2018.1493353.
68. Bruce Thyer, 'A Bibliography of Randomized Controlled Experiments in Social Work (1949–2013)', Research on Social Work Practice 25 no.7 (2015): 753-793.
69. Find out more about the Global Policing Database: https://global-policing-database.dhssp.cloud.edu.au/s/gpd/page/about
70. Robert G. St. Pierre and Jean I. Layzer, 'Using Home Visits for Multiple Purposes: The Comprehensive Child Development Program,' The Future of Children 9, no. 1, spring/summer (1999): 134-151; Robert G. St Pierre et al, National Impact Evaluation of the Comprehensive Child Development Program (Administration for Children and Families, U.S. Department of Health and Human Services, 1997) https://www.acf.hhs.gov/sites/default/files/opre/ccdp_fullreport.pdf
71. Find out more about Creative Credit at: https://www.nesta.org.uk/feature/measuring-our-impact/creative-credits
72. Cancer Research UK discusses the phases on clinical trials at: https://www.cancerresearchuk.org/about-cancer/find-a-clinical-trial/what-clinical-trials-are/phases-of-clinical-trials
73. Graham Moore et al, Process evaluation of complex interventions: UK Medical Research Council (MRC) guidance (MRC Population Health Science Research Network, 2015) https://mrc.ukri.org/documents/pdf/mrc-phsrn-process-evaluation-guidance-final
74. Angus Deaton and Nancy Cartwright, 'Understanding and misunderstanding randomized controlled trials', Social Science & Medicine 210 (2018): 2-21. https://www.sciencedirect.com/science/article/pii/S0277953617307359
75. For more details, see Nesta's Innovation Growth Lab Experimentation Toolkit, Section 1.7 on Limitations of RCTs, included in our list of further resources in Section 4. https://www.nesta.org.uk/toolkit/innovation-growth-lab-experimentation-toolkit
76. Dougal Hutchison and Ben Styles, A Guide to Running Randomised Controlled Trials for Educational Researchers. (NFER, Slough, 2010) https://www.nfer.ac.uk/media/2114/rct01.pdf
77. Angus Deaton and Nancy Cartwright, 'Understanding and misunderstanding randomized controlled trials' Social Science & Medicine 210 (2018): 2-21. https://www.sciencedirect.com/science/article/pii/S0277953617307359
78. See Mary Ann Bates and Rachel Glennerster, 'The Generalizability Puzzle', Stanford Social Innovation Review (Summer 2017). Available through Open Access online at: https://ssir.org/articles/entry/the_generalizability_puzzle#
79. Gerry Stoker, 'Exploring the Promise of Experimentation in Political Science: Micro-Foundational Insights and Policy Relevance', Political Studies 58 no. 2 (2010).: 300–319. https://doi.org/10.1111/j.1467-9248.2009.00812.x p303.
80. Jonathan Breckon, Better Public Services Through Experimental Government (Alliance for Useful Evidence, Nesta, London) p32.
81. Jonathan Breckon, Better Public Services Through Experimental Government (Alliance for Useful Evidence, Nesta, London) p43.
82. Phil Alderson, Ian Roberts, 'Corticosteroids for acute traumatic brain injury', Cochrane Database of Systematic Reviews 1 Art. No. CD000196 (2005). Available from Cochrane Library: https://www.cochrane.org/CD000196/INJ_corticosteroids-to-treat-brain-injury
83. For example, if an intervention is shown to be effective in one setting but uncertainty remains about how it will work in others, how much evidence – and how many experiments - do we need to make further trials unnecessary, or unethical?
84. See Demos Helsinki's website, the organisation helped author the Code of Conduct: https://www.demoshelsinki.fi/en/referenssit/code-conduct-national-ethical-code-conduct-societal-experiments
85. Adrian O'Dowd, 'Scientists call for more multi-arm clinical trials to speed up approval of new drugs', BMJ 349:g4812 (2014) https://www.bmj.com/content/349/bmj.g4812
86. MRC Blog, 'Clinical trials: why multi-arms are better than two', Medical Research Council Blog, 25 July 2014 https://mrc.ukri.org/news/blog/clinical-trials-why-multi-arms-are-better-than-two
87. Ibid.
88. See Optimizely's Glossary: https://www.optimizely.com/uk/optimization-glossary/multi-armed-bandit
89. Sofia Villar, Jack Bowden, James Wason, 'Multi-armed Bandit Models for the Optimal Design of Clinical Trials: Benefits and Challenges', Statistical Science 30 no. 2 (2015):199–215.4. Statistical Science 30(2):199–215.4.
90. Chris Kaibel and Torsten Biemann, 'Rethinking the Gold Standard With Multi-armed Bandits: Machine Learning Allocation Algorithms for Experiments', Organizational Research Methods (2019) Organizational Research Methods.
91. Aaron Dibner-Dunlap and Yumna Rathore, 'Beyond RCTs: How Rapid-Fire Testing Can Build Better Financial Products', IPA Blog, 2016 https://www.poverty-action.org/blog/beyond-rcts-how-rapid-fire-testing-can-build-better-financial-products
92. Find out more about the EdTech Innovation Testbed: https://www.nesta.org.uk/project/edtech-innovation-testbed

93. Alaka Holla, 'Seeking Nimble Plumbers', World Bank Blogs, April 4 2018 https://blogs.worldbank.org/impactevaluations/seeking-nimble-plumbers
94. Read a summary of the World Bank's Nimble Evaluations: https://www.worldbank.org/en/programs/sief-trust-fund/brief/nimble-summaries
95. Dean Karlan and Mary Kay Gugert, The Goldilocks Challenge: Right-Fit Evidence for the Social Sector (OUP USA, 2018).
96. Adapted from Matthew Salganik, Bit by Bit: Social Research in the Digital Age. (Princeton, NJ: Princeton University Press, 2017) Chapter 4: Running Experiments.
97. Behavioural Insights Team et al, Applying Behavioural Insights to Organ Donation: Preliminary results from a randomised controlled trial Cabinet Office, London, 2013).
98. Joseph Henrich, Steve J. Heine, Ara Norenzayan, 'The Weirdest People in the World?,' Working Paper Series of the German Council for Social and Economic Data 139, German Council for Social and Economic Data (RatSWD) (2010) https://ideas.repec.org/p/rsw/rswwps/rswwps139.html
99. Amazon Mechanical Turk (MTurk): https://www.mturk.com
100. LabintheWild: https://www.labinthewild.org
101. The Web Experiment List: http://www.wexlist.net
102. Damon Centola, How Behaviour Spreads (Princeton, NJ: Princeton University Press, 2018).
103. Adam Kramer, Jamie Guillory, Jeffrey Hancock, 'Experimental evidence of massive-scale emotional contagion through social networks', Proc Natl Acad Sci U S A 111 no. 24 (2014):8788–8790. doi: 10.1073/pnas.1320040111.
104. Michael H. Birnbaum, 'Human Research and Data Collection via the Internet', Annual Review of Psychology 55, no. 1 (2004): 803-832. https://doi.org/10.1146/annurev.psych.55.090902.141601
105. Jesse Chandler et al, 'Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers', Behavior Research Methods 46 (2013): 112. https://doi.org/10.3758/s13428-013-0365-7 doi: 10.3758/s13428-013-0365-7
106. Andrew Leigh, Randomistas: How Radical Researchers Changed Our World (Yale University Press, 2018).
107. Andrew Leigh, Randomistas: How Radical Researchers Changed Our World (Yale University Press, 2018) p131.
108. See Obama for America's Digital Fundraising Machine: https://www.optimizely.com/uk/customers/obama2012
109. Behavioural Insights Team et al, Applying Behavioural Insights to Organ Donation: preliminary results from a randomised controlled trial (Behavioural Insights Team, Cabinet Office, London, 2013) https://www.bi.team/publications/applying-behavioural-insights-to-organ-donation
110. Ibid.
111. Matthew Salganik, Bit by Bit: Social Research in the Digital Age (Princeton, NJ: Princeton University Press, 2017).
112. Test & Build platform: https://www.testandbuild.com; Predictiv platform: https://www.bi.team/bi-ventures/predictiv
113. POST, Understanding Research Methods (Parliamentary Office for Science and Technology, forthcoming in 2020).
114. Thomas Dee and Benjamin Keys, 'Does merit pay reward good teachers? Evidence from a randomized experiment', Journal of Policy Analysis and Management 23 no. 3 (Summer 2004): 471–488.
115. For example, in a subsidised meal programme experiment in Kenya, upset parents in over half of the control schools organised to raise funds for student meals to match what was being received in the treatment group. See David Greenber and Mark Shroder, The Digest of Social Experiments (3rd ed.) (Washington, DC: Urban Institute Press, 2004) p.399.
116. Gary King, Emmanuela Gakidou, Nirmala Ravishankar, Ryan Moore, Jason Lakin, Manett Vargas et al, 'A 'politically robust' experimental design for public policy evaluation, with application to the Mexican universal health insurance program', Journal of Policy Analysis and Management 26 (2007): 479-506.
117. Ibid.
118. Laura Haynes, Owain Service, Ben Goldacre, David Torgerson, Test, Learn, Adapt: Developing Public Policy with Randomised Controlled Trials (Cabinet Office, London, 2010) p.22. https://www.gov.uk/government/publications/test-learn-adapt-developing-public-policy-with-randomised-controlled-trials
119. Emma Beard et al, 'Stepped wedge randomised controlled trials: systematic review of studies published between 2010 and 2014', Trials 16, no. 353 (2015).
120. Noreen Mdege et al, 'Systematic review of stepped wedge cluster randomized trials shows that design is particularly used to evaluate interventions during routine implementation' Journal of Clinical Epidemiology Sep;64(9) (2011): 936-48. doi: 10.1016/j.jclinepi.2010.12.003
121. Dominic Pearson, David Torgerson, Cynthia McDougall, Roger Bowles 'A parable of two agencies, one of which randomises', Annals of the American Academy of Political & Social Sciences 628 (2010): 11-29.
122. Ministry of Housing, Communities and Local Government, Measuring the impact of Community-Based English Language Provision Findings from a Randomised Controlled Trial (National Learning and Work Institute, Ministry of Housing, Communities and Local Government, HM Government, 2018).

123. Interview with Dr Jen Gold, Head of the What Works Team.

124. Ministry for Housing, Communities and Local Government, Integrated Communities Strategy green paper (Ministry for Housing, Communities and Local Government, HM Government, 2018).

125. What Works Network, The Rise of Experimental Government: Cross Government Trials Advice Panel Update Report, (What Works Network and Economic and Social Research Council, London) p17.

126. Neil Salkind, 'Crossover Design', defined in SAGE Encyclopaedia of Research Design, SAGE Publishing, 2010 https://methods.sagepub.com/reference/encyc-of-research-design/n95.xml

127. Kerry Dwan, Tianjing Li, Douglas G Altman, Diana Elbourne, 'CONSORT 2010 statement: extension to randomised crossover trials', BMJ 366 (2019) :l4378. https://www.bmj.com/content/366/bmj.l4378

128. Hui-Ling Lai and Yin-Ming Li, 'The effect of music on biochemical markers and self-perceived stress among first-line nurses: a randomized controlled crossover trial,' Journal of Advanced Nursing 67(11) (2011): 2414-24. https://doi.org/10.1111/j.1365-2648.2011.05670.x

129. Halima Khan, 'Nesta's work on Personalised Care: How we got here', Nesta Blog, 4 February 2019 https://www.nesta.org.uk/blog/nestas-work-personalised-care-how-we-got-here

130. See Health Education England's page on Person-centred Care: https://www.hee.nhs.uk/our-work/person-centred-care

131. Eva van der Ploeg, Barbara Eppingstall, Cameron Camp, Susannah Runci, John Taffe, Daniel O'Connor, 'A randomized crossover trial to study the effect of personalized, one-to-one interaction using Montessori-based activities on agitation, affect, and engagement in nursing home residents with Dementia', International Psychogeriatrics 25(4), (2013): 565-575. https://doi.org/10.1017/S1041610212002128

132. Kerry Dwan, Tianjing Li, Douglas G Altman, Diana Elbourne, 'CONSORT 2010 statement: extension to randomised crossover trials', BMJ 366 (2019) :l4378. https://www.bmj.com/content/366/bmj.l4378

133. Dean Karlan and Jacob Appe, Failing in the Field (Princeton, NJ: Princeton University Press, 2016) p8.

134. University of Cambridge Research News: https://www.cam.ac.uk/research/news/first-scientific-report-shows-police-body-worn-cameras-can-prevent-unacceptable-use-of-force

135. Alex Sutherland et al, 'Post-experimental follow-ups – Fade-out versus persistence effects: The Rialto police body-worn camera experiment four years on', Journal of Criminal Justice 53 (November 2017): 110-116.

136. Barak Ariel, et al, 'The Effect of Police Body-Worn Cameras on Use of Force and Citizens' Complaints Against the Police: A Randomized Controlled Trial', Journal of Quantitative Criminology 31 no.3 (November 2014): 509-535.

137. Barak Ariel et al, 'Report: increases in police use of force in the presence of body-worn cameras are driven by officer discretion: a protocol-based subgroup analysis of ten randomized experiments', Journal of Experimental Criminology 12 (2016): 453-463; Barak Ariel et al, 'Contagious Accountability': A Global Multisite Randomized Controlled Trial on the Effect of Police Body-Worn Cameras on Citizens' Complaints Against the Police', Criminal Justice and Behavior, 44(2) (2017): 293–316. https://doi.org/10.1177/0093854816668218

138. Alex Sutherland et al, 'Post-experimental follow-ups – Fade-out versus persistence effects: The Rialto police body-worn camera experiment four years on', Journal of Criminal Justice 53 (November 2017): 110-116.

139. This has been discussed by researchers over the years. See this perspective from the US in 2004: Lawrence Sherman and Heather Strang, 'Experimental Ethnography: The Marriage of Qualitative and Quantitative Research', The ANNALS of the American Academy of Political and Social Science 595 no. 1 (2004): 204–222. https://doi.org/10.1177/0002716204267481

140. Ray Pawson and Nicholas Tilley, Realistic Evaluation (London: SAGE Publications Ltd, 1997, 2014); Chris Bonnell et al, 'Realist randomised controlled trials: A new approach to evaluating complex public health interventions', Social Science & Medicine 75(12) (December 2012): 2299-306. doi: 10.1016/j.socscimed.2012.08.032

141. Peter Hedström and Petri Ylikoski, 'Causal Mechanisms in the Social Sciences', Annual Review of Sociology Vol 36 (2010): 49-67. Doi: 10.1146/annurev.soc.012809.102632

142. Ray Pawson and Nicholas Tilley, Realistic Evaluation (London: SAGE Publications Ltd, 1997, 2014)

143. Farah Jamal et al, 'The three stages of building and testing mid-level theories in a realist RCT: A theoretical and methodological case-example', Trials 16, no. 466 (2015) Open Access: https://trialsjournal.biomedcentral.com/articles/10.1186/s13063-015-0980-y

144. Chris Bonnell et al, 'Realist randomised controlled trials: A new approach to evaluating complex public health interventions', Social Science & Medicine 75(12) (December 2012): 2299-306. doi: 10.1016/j.socscimed.2012.08.032

145. Lyndal Bond et al, 'The Gatehouse Project: Can a multilevel school intervention affect emotional wellbeing and health risk behaviours?', Journal of Epidemiology and Community Health 58, no.12 (2004): 997–1003. doi: 10.1136/jech.2003.009449

146. Wolfgang Markham and Paul Aveyard, 'A new theory of health promoting schools based on human functioning, school organisation and pedagogic practice', Social Science & Medicine 56, no.6 (2003): 1209-20. doi: 10.1016/s0277-9536(02)00120-x

147. Chris Bonell, et al, 'Initiating change locally in bullying and aggression through the school environment (INCLUSIVE): a pilot randomised controlled trial', Health Technology Assessment 19(53) (2015).

148. Chris Bonell et al, 'Examining intervention mechanisms of action using mediation analysis within a randomised trial of a whole-school health intervention', Journal of Epidemiology & Community Health 73 no. 5 (2019): 455-464.

149. LSHTM, 'Making UK schools more inclusive places could help reduce bullying and promote well-being', London School of Hygiene & Tropical Medicine Blog, 2018 https://www.lshtm.ac.uk/newsevents/news/2018/making-uk-schools-more-inclusive-places-could-help-reduce-bullying-and-promote

150. Chris Bonell et al, 'Examining intervention mechanisms of action using mediation analysis within a randomised trial of a whole-school health intervention', Journal of Epidemiology & Community Health 73 no. 5 (2019): 455-464.

151. C. Hendricks Brown et al, 'An Overview of Research and Evaluation Designs for Dissemination and Implementation', Annual Review of Public Health 38 (March 2017): 1-22.

152. Ibid.

153. Geoffrey Curran et al, 'Effectiveness-Implementation Hybrid Designs', Medical Care 50, no.3 (March 2012): 217-26.

154. C. Hendricks Brown et al, 'An Overview of Research and Evaluation Designs for Dissemination and Implementation', Annual Review of Public Health 38 (March 2017): 1-22.

155. Geoffrey Curran et al, 'Implementation of the CALM intervention for anxiety disorders: A qualitative study', Implementation Science 7, no. 14 (2012) doi:10.1186/1748-5908-7-14. Open Access Review: https://implementationscience.biomedcentral.com/articles/10.1186/1748-5908-7-14

156. Bryan Garner et al, 'Testing the effectiveness of a motivational interviewing-based brief intervention for substance use as an adjunct to usual care in community-based AIDS service organizations: Study protocol for a multisite randomized controlled trial', Addiction Science & Clinical Practice, 12, no.31 (2017) doi: 10.1186/s13722-017-0095-8

157. Kari Gloppen et al, 'Sustaining adoption of science-based prevention through communities that care', Journal of Community Psychology 44:1 (2016) https://onlinelibrary.wiley.com/doi/abs/10.1002/jcop.21743

158. Registration of The Community Youth Development Study on ClinicalTrials.gov: https://clinicaltrials.gov/ct2/show/study/NCT01088542?show_desc=Y#desc

159. Weichung Joe Shih, 'Plan to be Flexible: A Commentary on Adaptive Designs', Biometrical Journal / Biometrische Zeitschrift 48:4 (2006) 656-9; discussion 660. doi: 10.1002/bimj.200610241.

160. Philip Pallmann et al, 'Adaptive designs in clinical trials: Why use them, and how to run and report them', BMC medicine 16(1), no. 29 (2018) doi:10.1186/s12916-018-1017-7.

161. Ibid.

162. Margaret Handley et al, 'Selecting and Improving Quasi-Experimental Designs in Effectiveness and Implementation Research', Annual Review of Public Health 39:1 (2018): 5-25. doi: 10.1146/annurev-publhealth-040617-014128.

163. Sanjay Basu, Ankita Meghani, Arjumand Siddiqi, 'Evaluating the Health Impact of Large-Scale Public Policy Changes: Classical and Novel Approaches', Annual Review of Public Health 38 (2017): 351–370. doi: 10.1146/annurev-publhealth-031816-044208.

164. Steven Glazerman,Dan Levy, David Myers, 'Nonexperimental versus Experimental Estimates of Earnings Impacts', The Annals of the American Academy of Political and Social Science 589 (2003): 63-93. Retrieved from www.jstor.org/stable/3658561

165. Sanjay Basu, Ankita Meghani, Arjumand Siddiqi, 'Evaluating the Health Impact of Large-Scale Public Policy Changes: Classical and Novel Approaches', Annual Review of Public Health 38 (2017): 351–370. doi: 10.1146/annurev-publhealth-031816-044208.

166. Martin Ravallion, Should the Randomistas (Continue to) rule? (Working Paper 492, Centre for Global Development, August 2018) https://www.cgdev.org/publication/should-randomistas-continue-rule

167. Margaret Handley et al, 'Selecting and Improving Quasi-Experimental Designs in Effectiveness and Implementation Research', Annual Review of Public Health 39:1 (2018): 5-25. doi: 10.1146/annurev-publhealth-040617-014128; Sanjay Basu, Ankita Meghani, Arjumand Siddiqi, 'Evaluating the Health Impact of Large-Scale Public Policy Changes: Classical and Novel Approaches', Annual Review of Public Health 38 (2017): 351–370. doi: 10.1146/annurev-publhealth-031816-044208.

168. Thomas Cook, ''Waiting for Life to Arrive': A history of the regression-discontinuity design in Psychology, Statistics and Economics', Journal of Econometrics 142, no. 2, (February 2008): 636-654.

169. Ibid.
170. Paul Gertler et al, Impact Evaluation in Practice (Washington DC: The International Bank for Reconstruction & Development/World Bank, 2011).
171. Ibid.
172. Ibid.
173. Adapted from Paul Gertler et al, Impact Evaluation in Practice (Washington DC: The International Bank for Reconstruction & Development / World Bank, 2011) p103.
174. Ojmarrh Mitchell, Joshua Cochran, Daniel Mears, William Bales, 'Examining Prison Effects on Recidivism: A Regression Discontinuity Approach', Justice Quarterly 34:4 (2017): 571-596. doi: 10.1080/07418825.2016.1219762.
175. European Commission, Testing Social Policy Innovation (Employment, Social Affairs & Inclusion, European Commission, 2014).
176. Ibid.
177. William Shadish, Thomas Cook, and Donald Campbell, Experimental and quasi-experimental designs for generalized causal inference (Wadsworth Cengage Learning, Boston, MA, 2002).
178. David Lee and Thomas Lemieux (2010) Regression Discontinuity Designs in Economics Journal of Economic Literature 48 281–355. Available online: https://www.princeton.edu/~davidlee/wp/RDDEconomics.pdf
179. Paul Gertler et al, Impact Evaluation in Practice (Washington DC: The International Bank for Reconstruction & Development/World Bank, 2011).
180. Sanjay Basu, Ankita Meghani, Arjumand Siddiqi, 'Evaluating the Health Impact of Large-Scale Public Policy Changes: Classical and Novel Approaches', Annual Review of Public Health 38 (2017): 351–370. doi: 10.1146/annurev-publhealth-031816-044208.
181. Paul Gertler et al, Impact Evaluation in Practice (Washington DC: The International Bank for Reconstruction & Development/World Bank, 2011).
182. Martin Ravallion, Should the Randomistas (Continue to) rule? (Working Paper 492, Centre for Global Development, August 2018) https://www.cgdev.org/publication/should-randomistas-continue-rule
183. All evaluation documents are available at the evaluation homepage at Birkbeck University's site: http://www.bbk.ac.uk
184. Sarah Cattan et al, The health effects of Sure Start (Institute for Fiscal Studies, London, June 2019) https://www.ifs.org.uk/publications/14139
185. Find out more about the Millennium Cohort Study: cls.ucl.ac.uk/cls-studies/millennium-cohort-study
186. Professor Edward Melhuish et al, The Impact of Sure Start Local Programmes on Seven Year Olds and Their Families, (The National Evaluation of Sure Start (NESS) Team, Institute for the Study of Children, Families and Social Issues; Birkbeck, University of London; and the Department for Education, 2012).
187. Ibid.
188. Ibid.
189. Sarah Cattan et al, The health effects of Sure Start (Institute for Fiscal Studies, London, June 2019) https://www.ifs.org.uk/publications/14139
190. Paul Gertler et al, Impact Evaluation in Practice (Washington DC: The International Bank for Reconstruction & Development/World Bank, 2011)
191. Ibid.
192. Margaret Handley et al, 'Selecting and Improving Quasi-Experimental Designs in Effectiveness and Implementation Research', Annual Review of Public Health 39:1 (2018): 5-25. doi: 10.1146/annurev-publhealth-040617-014128.
193. Coady Wing et al, 'Designing Difference in Difference Studies: Best Practices for Public Health Policy Research', Annual Review of Public Health 39 (2018): 453-469; John Snow, On the Mode of Communication of Cholera London (John Churchill, London, 1855).
194. Simon Burgess, Deborah Wilson, Jack Worth, A natural experiment in school accountability: The impact of school performance information on pupil progress and sorting (The Centre For Market And Public Organisation, Bristol University, Brisol, October 2010) http://www.bristol.ac.uk/media-library/sites/cmpo/migrated/documents/wp246.pdf p1.
195. Ibid
196. Coady Wing et al, 'Designing Difference in Difference Studies: Best Practices for Public Health Policy Research', Annual Review of Public Health 39 (2018): 453-469.
197. Ibid.
198. Sanjay Basu, Ankita Meghani, Arjumand Siddiqi, 'Evaluating the Health Impact of Large-Scale Public Policy Changes: Classical and Novel Approaches', Annual Review of Public Health 38 (2017): 351–370. doi: 10.1146/annurev-publhealth-031816-044208.
199. Margaret Handley et al, 'Selecting and Improving Quasi-Experimental Designs in Effectiveness and Implementation Research', Annual Review of Public Health 39:1 (2018): 5-25. doi: 10.1146/annurev-publhealth-040617-014128.
200. Ibid.

201. Coady Wing et al, 'Designing Difference in Difference Studies: Best Practices for Public Health Policy Research', Annual Review of Public Health 39 (2018): 453-469.
202. Note that averages are 'weighted', which affects how they are combined. See Robert McClelland and Sarah Gault, The Synthetic Control Method as a Tool to Understand State Policy' (Urban Institute, Washington DC, March 2017).
203. Alberto Abadie and Javier Gardeazabal, 'The Economic Costs of Conflict: A Case Study of the Basque Country', American Economic Review 93, no. 1 (March 2003): 113-132. doi: 10.1257/000282803321455188.
204. Robert McClelland and Sarah Gault, The Synthetic Control Method as a Tool to Understand State Policy' (Urban Institute, Washington DC, March 2017).
205. Ibid.
206. Alberto Abadie, Alexis Diamond, Jens Hainmueller, 'Synthetic Control Methods for Comparative Case Studies', Journal of the American Statistical Association 105, No. 490, Applications and Case Studies (2010).
207. Sanjay Basu, Ankita Meghani, Arjumand Siddiqi, 'Evaluating the Health Impact of Large-Scale Public Policy Changes: Classical and Novel Approaches', Annual Review of Public Health 38 (2017): 351–370. doi: 10.1146/annurev-publhealth-031816-044208.
208. Ibid.
209. Michihito Ando, 'Dreams of Urbanization: Quantitative Case Studies on the Local Impacts of Nuclear Power Facilities Using the Synthetic Control Method', Journal of Urban Economics 85 (January 2015): 68-85.
210. Robert McClelland and Sarah Gault, The Synthetic Control Method as a Tool to Understand State Policy' (Urban Institute, Washington DC, March 2017).
211. Ibid.
212. Adapted from Hank Jenkins-Smith et al, Quantitative Research Methods for Political Science, Public Policy and Public Administration: 4th Edition with Applications in R (2017), a free resource available online from multiple sources. See: https://bookdown.org/ripberjt/qrmbook
213. Julie Reed, Alan Card, 'The problem with Plan-Do-Study-Act cycles', BMJ Quality & Safety 25 (2016):147-152; Ann Mei Chang, Lean Impact: How to Innovate for Radically Greater Social Good (John Wiley & Sons, 2018).
214. You can find out more on Dartington's approach and its work with Chance UK, a mentoring charity, on its website in this blog article: Deon Simpson, 'To help London's children, one mentoring service is improving itself first', Dartingon Service Design Lab blog, 26 July 2019 https://www.dartington.org.uk/ourblog/tohelplondonchildren
215. Find out more at: https://developingchild.harvard.edu/science/national-scientific-council-on-the-developing-child
216. Maia Barrow et al, 'Improved cognitive flexibility after a structured play intervention with a high-risk sample of preschoolers', (2015) doi: 10.13140/RG.2.1.2093.2966.
217. Find out more online about People Powered Results: https://www.nesta.org.uk/project/people-powered-results
218. Interview with Brad Dudding, formerly Chief Impact Officer at the Centre for Employment Opportunities.
219. Dartington Service Design Lab and Family Nurse Partnership National Unit, FNP Adapt Interim Report (Interim report of the FNP ADAPT project, London, 2018). http://fnp.nhs.uk/media/1246/fnp-adapt-interim-report.pdf
220. Tom McBride, 'What Happens When the Evidence Is Mixed?' Early Intervention Foundation blog, 29 March 2018 https://www.eif.org.uk/blog/what-happens-when-the-evidence-is-mixed
221. Michael Robling et al, 'Effectiveness of a Nurse-Led Intensive Home-Visitation Programme for First-Time Teenage Mothers (Building Blocks): A Pragmatic Randomised Controlled Trial,' The Lancet 387, no. 10014 (2016): 146–55. See also the process evaluation: J. Sanders et al, 'Implementation of the Family Nurse Partnership programme in England: Experiences of key health professionals explored through trial parallel process evaluation', BMC Nursing 18, no. 13 (2019) doi:10.1186/s12912-019-0338-y Available Open Access: https://bmcnurs.biomedcentral.com/articles/10.1186/s12912-019-0338-y
222. See 'FNP Next Steps: Developing and Improving the Programme', Family Nurse Partnership: https://fnp.nhs.uk/fnp-next-steps
223. Dr Nick Axford, 'Applying "Dark Logic" in Practice', Alliance for Useful Evidence blog, 29 November 2017 https://www.alliance4usefulevidence.org/applying-dark-logic-in-practice
224. Dartington Service Design Lab and Family Nurse Partnership National Unit, FNP Adapt Interim Report (Interim report of the FNP ADAPT project, London, 2018). http://fnp.nhs.uk/media/1246/fnp-adapt-interim-report.pdf
225. Interview with Tim Hobbs, Director of Dartington Service Design Lab.
226. Interview with Jon Brown, Head of Impact at NSPCC.
227. See Geoff Mulgan, 'Teach a Society to Fish: The twin revolutions that could transform social science', Nesta blog, 22 September 2017 https://www.nesta.org.uk/blog/teach-a-society-to-fish-the-twin-revolutions-that-could-transform-social-science; and Geoff Mulgan, 'Social Science and Intelligence Design', Nesta blog, 25 January 2019 https://www.nesta.org.uk/blog/social-science-and-intelligence-design
228. See our list of useful resources in Section 4.

229. Nesta, 'Prototyping' in 20 Tools for Innovation in Government (Nesta, London, September 2019) https://www.nesta.org.uk/report/20-tools-innovating-government

230. Nesta, Prototyping in Public Services (Nesta, London, November 2011) https://www.nesta.org.uk/report/prototyping-in-public-services

231. Bella Martin and Bruce Hanington, Universal Methods of Design: 100 Ways to Research Complex Problems, Develop Innovative Ideas, and Design Effective Solutions (Gloucester, MA: Rockport Publishers Inc, February 2012); Camilla Buchanan, 'Prototyping for Policy', Open Policy blog, 27 November 2018 https://openpolicy.blog.gov.uk/2018/11/27/prototyping-for-policy

232. For guides on roleplaying, creating simulations and other ideas, see our Designing for Public Services Guide in Section 4: Useful resources.

233. Camilla Buchanan, 'Prototyping for Policy', Open Policy blog, 27 November 2018 https://openpolicy.blog.gov.uk/2018/11/27/prototyping-for-policy

234. GOV.UK Prototype Kit: https://govuk-prototype-kit.herokuapp.com/docs

235. James K Esser, 'Alive and Well after 25 Years: A Review of Groupthink Research, Organizational Behavior and Human Decision Processes', 73, no. 2–3 (1998): 116-141. ISSN 0749-5978 https://doi.org/10.1006/obhd.1998.2758

236. Ian Sanderson, 'Evaluation, Policy Learning and Evidence-Based Policy Making', Public Administration 80, no. 1 (2002) 1-22. doi: 10.1111/1467-9299.00292.

237. Bella Martin and Bruce Hanington, Universal Methods of Design: 100 Ways to Research Complex Problems, Develop Innovative Ideas, and Design Effective Solutions (Gloucester, MA: Rockport Publishers Inc, February 2012).

238. Bruce Hanington, 'Relevant and Rigorous: Human-Centered Research and Design Education', Design Issues 26, no. 3 (Summer 2010): 18-26. Retrieved from: http://www.jstor.org/stable/20749956

239. Peter H. Rossi, 'The Iron Law of Evaluation and Other Metallic Rules'. Research in Social Problems and Public Policy 4 (1987): 3-20.

240. Andrew Leigh, Randomistas: How Radical Researchers Changed Our World (Yale University Press, 2018).

241. Kathryn Oliver and Annette Boaz, 'Transforming evidence for policy and practice: Creating space for new conversations', Palgrave Commununications 5, no. 60 (2019) doi: 10.1057/s41599-019-0266-1 Open Access at: https://www.nature.com/articles/s41599-019-0266-1

242. Nesta, 'The UK What Works Network' in 20 Tools for Innovation in Government (Nesta, London, September 2019) https://www.nesta.org.uk/report/20-tools-innovating-government

243. See Jill Rutter, 'Troubled Familiies Takeaways', Institute for Government blog, 18 October 2016 https://www.instituteforgovernment.org.uk/blog/troubled-families-takeaways

244. See the 2019 evaluation by Ministry of Housing, Communities and Local Government, National evaluation of the Troubled Families Programme 2015-2020: Findings (Ministry of Housing, Communities and Local Government, 2019).

245. Cambridge Online Dictionary definition of innovation at time of writing.

246. See European Commission website on Social Innovation: https://ec.europa.eu/growth/industry/innovation/policy/social_en

247. Judea Pearl and Elias Bareinboim, Transportability of Causal and Statistical Relations: A Formal Approach, (Proceedings of the 25th AAAI Conference on Artificial Intelligence, August 7–11, 2011, San Francisco, CA).

248. Michelle Jackson and D.R. Cox, 'The Principles of Experimental Design and Their Application in Sociology', Annual Review of Sociology 39 (2013): 27-49. Retrieved from http://www.jstor.org/stable/43049624

249. Matthew Salganik, Bit by Bit: Social Research in the Digital Age (Princeton, NJ: Princeton University Press, 2017) p176.

250. Ben Styles and Dougal Hutchinson, A Guide to Running Randomised Controlled Trials for Educational Researchers (Slough: NFER, 2010).

251. Kelsey Piper, 'The Nobel went to economists who changed how we help the poor. But some critics oppose their big idea', Vox, 11 December 2019 https://www.vox.com/future-perfect/2019/12/11/20938915/nobel-prize-economics-banerjee-duflo-kremer-rcts

About Nesta

Nesta is an innovation foundation. For us, innovation means turning bold ideas into reality and changing lives for the better.

We use our expertise, skills and funding in areas where there are big challenges facing society.

Nesta is based in the UK and supported by a financial endowment. We work with partners around the globe to bring bold ideas to life to change the world for good.

www.nesta.org.uk

58 Victoria Embankment
London EC4Y 0DS

+44 (0)20 7438 2500
information@nesta.org.uk
@nesta_uk
www.facebook.com/nesta.uk
www.nesta.org.uk