Funded by UK Government
Centre for Homelessness Impact
SCHOOL FOR GOVERNMENT
KING'S College LONDON

Designing and delivering randomised trials for social policy

Michael Sanders | Julia Ellingwood | Dimitris Vallis |
Vanessa Hirneis | Kira Ewanich | Isobel Harrop |
Guillermo Rodriguez | Susannah Hume | Eliza Kozman |

February 2025

Authors

Michael Sanders
Director, School for Government at King's College London

Dimitris Vallis
Research Associate, the Policy Institute at King's College London

Kira Ewanich
Research Assistant, the Experimental Government Team at King's College London

Eliza Kozman
Deputy Chief Executive Officer TASO

Susannah Hume
Director of Evaluation, the Policy Institute at King's College London

Julia Ellingwood
Research Associate, the Experimental Government Team at King's College London

Isobel Harrop
Research Assistant, the Experimental Government Team at King's College London

Guillermo Rodriguez
Director of Evidence and Data , Centre for Homelessness Impact

Vanessa Hirneis
Social Psychologist and Behavioural Researcher, the Experimental Government Team at King's College London

Acknowledgements

This project was funded by a grant from the Evaluation Task Force (Cabinet Office/HM Treasury) and its Evaluation Accelerator Fund (2023-2025).

We are also indebted to the Centre for Homelessness Impact (CHI), who committed funding and provided extensive expertise based on their experiences running and supporting trials to reduce homelessness.

Transforming Access and Student Outcomes in Higher Education (TASO) were also instrumental in the development of this book, lending expertise in evidence generation in higher education.

Permission to share

Contents

Foreword

Professor David Halpern

> I had considerable freedom of choice of therapy: my trouble was that I did not know which to use and when. I would gladly have sacrificed my freedom for a little knowledge. I had never heard of 'randomized controlled trials', but I knew that there was no real evidence that anything we had to offer had any effect on tuberculosis, and I was afraid that I shortened the lives of some of my friends by unnecessary intervention.

It is just over 50 years since Archie Cochrane published his measured but passionate book on Efficiency and Effectiveness (1972), and little over 25 years since the creation of National Institute for Health and Care Excellence (NICE), the first institution to take that evidence and systematically use it to shape healthcare policy and practice.

Before Cochrane, systematically running randomised controlled trials to evaluate the impact of medicines was not just unheard of, it was unthinkable – and indeed, Cochrane faced significant opposition from the medical establishment to adopting the scientific approach to treatment.

Over the last 15 years, there has been a rapid expansion in the application of experimental methods and RCTs to other areas of policy and practice. That is good news. But the bad news is that we have far, far to go. The number of RCTs conducted in health is around 10-fold of those conducted in every other field combined.

The expansion – particularly in the UK – has been characterised by the creation of a number of what works centres. Many of these have gone beyond the synthesis remit of NICE, to actually commissioning and conducting trials themselves – shining an quantitative empirical light into areas they had not previously been considered – from children's social care, to homelessness, to ageing. Government has invested not only in these centres, but in creating the Evaluation Taskforce, and aspiring to a higher standard of evaluation across government.

People often ask: why have experimental methods been so slow to be adopted? There are many barriers. Political and sometimes even constitutional can block the way, a worthy aspiration to treat everyone equally – even if we have no idea if it helps or hinders – trumping finding out what works. Cognitive or psychological barriers loom large too. Humans skew strongly to seeking confirming evidence over considering the counterfactual; to overconfidence; and to an 'illusion of explanatory depth' – things that are familiar are presumed to be understood.

Yet one of the biggest barriers, is often that public servants, professionals and governments lack the skills and familiarity of how to run a well-constructed RCT, let alone to build it into the routine of service delivery or practice. Although the central idea of a randomised controlled trial is simple, the potential complexity and variety is vast. Our paper, back in 2012, laid out the simple steps to running trials, but only went so far – acting as much as an advertisement for trials as a guide to conducting them.

I'm therefore delighted to see this new publication *Designing and delivering randomised trials for social policy* which goes into much more technical detail, on how to commission and conduct randomised controlled trials, in all their varieties. This new book will hopefully become an indispensable tool for civil servants, researchers and evaluators alike to help boost the quality of evidence, and through this, the quality of public services.

In a postscript to his famous book, Cochrane wrote: 'I may have been too critical of my colleagues for whom I actually have the greatest admiration and affection...What other profession encourages publications about its error, and experimental investigations into the effects of their actions? Which magistrate, judge, or headmaster has encouraged RCTs into their 'therapeutic' and 'deterrent' actions?'

I think Cochrane would have been delighted to see that there are now headmasters, and even a few judges, Ministers, police officers, and social workers encouraging such investigations. Yet we still have far to go.

David Halpern is a Visiting Professor in the School for Government at King's College London and President Emeritus at the Behavioural Insights Team. He was previously the government's first What Works National Advisor, from 2013 to 2022.

Preface

Each year, the government undertakes or pays for an enormous variety of activities – from national defence to benefits, from midwifery and early years education to pensions and palliative care. At the same time, government has always, and will always, experience far more demands for spending than there is money to be spent. In this context, it is crucial that money is spent as efficiently as possible, and that it achieves its ends as effectively as possible.

Alongside financial constraints, there are also human constraints across social policy. There are only so many social workers, and hence only so much social work that can be done. Children only have so many hours and years in school, and so we must choose what they are taught, from an almost infinite list of things that they could be taught.

In a democracy, deciding the goals of the public sector – whether we want to prioritise young people's learning of maths, or a broad curriculum including arts and theatre – is a matter for elected ministers or, ultimately, for parliament. However, once these priorities have been established, it is essential to understand which of the myriad different ways we could approach a problem will have the greatest impact and represent the best value for money.

Britain led the world in the late 1990s when it created the National Institute for Clinical Excellence (now the National Institute for Health and Care Excellence), or NICE, whose job is to decide what medicines are cost-effective and therefore suitable within the National Health Service.

But this revolution in cost-effective healthcare was only possible because of the revolution in medicine, starting after the second world war, that saw the use of randomised controlled trials become mainstream, and then essential, in producing drugs. The last decade and a half has seen successors to NICE, in the What Works Network, start up across many policy domains and by and large recognise that their circumstances are less auspicious than those faced by NICE, and that there are few high-quality randomised trials in their domains of interest. Many of these centres invested heavily in commissioning or conducting randomised controlled trials, helping to bridge the gap. Alongside this, government itself has become more engaged in running impact evaluations, thanks in no small part to the creation of the Evaluation Task Force which sits between the Cabinet Office and the Treasury.

Despite this investment, we are still in the foothills of an evidence-based policy revolution, one that depends on our ability to conduct randomised trials of high quality and at pace. This means learning from other fields, from our own experience, and from each other.

That is where the Ochre Book comes in. It aims to distil the basics of what kinds of trials exist and how to conduct them into a single document, to provide guidance to commissioners and evaluators alike that will help them understand better how trials can be run, and to help share our collective knowledge. Most importantly perhaps, we aim to reduce variation between trials, and to reduce the cost of running each trial.

Over the next several chapters, we will describe the various different types of trial, as well as insights on how to run them, and challenges in doing so. But before we start, this introduction covers the answers to some of the most basic questions – what is a trial, and why should we run one?

What is a randomised controlled trial?

A randomised controlled trial (RCT) is a research methodology that aims to find out whether something (an intervention), has an effect on something else (an outcome). It does this by recruiting a group of people who are eligible for the intervention, and then randomly choosing some proportion of them (typically, but not always, 50 per cent), to receive the intervention.

Because assignment is at random, it means that, in a large enough sample, people with different characteristics should be equally likely to end up in the treatment group or the control group. This applies for things we can easily observe – like gender, so there should be the same proportion of men in the treatment group as there is in the control group – and things that we can't – like motivation, so the average motivation of the two groups should also be the same.

If the two groups are the same on average on all these characteristics, then in the absence of any intervention, we'd expect them to have the same outcomes. This means the only difference between the two groups is that one received the intervention and the other didn't. Hence, any differences in outcomes we observe between the two groups at the end of the trial can be reasonably attributed to the effect of the intervention.

Why care about what works?

Why should we invest serious time and energy to finding out 'what works', which is the main question answerable by an RCT? The simple answer is that we don't already know what works in so many situations and across many policy areas. Time and again, studies have shown that interventions that were popular, or widely regarded as effective, are either ineffective, or actively harmful. Whether that's the Scared Straight intervention, which increased reoffending by juveniles; Achievement for All, which actively decreased educational attainment for children; or the Social Workers in School programme, which had no effect on social care outcomes, we are often surprised by these findings. Even if we could accurately predict the direction of an effect (which is a big if), then it is clearly harder to predict the magnitude of an effect, and hence, which of two effective interventions is the most cost-effective, and so should be chosen if we face a trade-off either in money or time.

Why run a trial?

The main reason to run an RCT is that they are the most reliable way to identify the impact of an intervention. Whether this is a form of financial support, a new way of doing teaching or social work, a form of support for your teams, or a training programme, an RCT can give you the cleanest sense of whether the intervention has an effect.

This is the case because we can't simply compare the people who choose to take up an intervention and compare them to people who didn't, and say that any difference is a result

of the treatment. Let's take an example from a trial we ran with the College of Policing that tested the effect of a new training programme on the use of force by the police. If we'd advertised the training to all police officers, and compared those who voluntarily took it with those that didn't, this measure of effectiveness could be biased, because the people who took up the training might be more enthusiastic about reducing force than those who didn't, and so would already be behaving differently, even in the absence of the intervention.

Similarly, we can't compare people's outcome before and after an intervention, because other things might have changed in the meantime. This is easy to see in an education context. Let's imagine we want to evaluate the impact of extra, after-school tuition, on maths performance, and we test participants' maths performance at the beginning of the school year and again at the end. The change in the scores over that time is caused by a soup of different factors – normal in-class teaching, homework, any self-study or support from parents, as well as the extra tuition. It's not possible to attribute any of the change to the tuition intervention specifically.

There are, of course, other ways of trying to identify the effects of interventions. These are collectively known as 'quasi-experimental approaches', and they can give good estimates of the effect of an intervention. However, they are mostly carried out retrospectively, and each method hinges on a particular set of assumptions – which might not be valid in a particular circumstance, and some of which cannot be validated until after the fact. These assumptions are more onerous, more difficult to validate, and less often met than the assumption on which an RCT rests – that randomisation has been successful. Importantly, they are also more 'data hungry' than an RCT, meaning that richer datasets and larger samples are needed to confidently detect effects. This might be fine for a popular intervention in a data-rich field like education, but is going to be harder for more targeted interventions, or in fields like housing and homelessness, where good- quality data is hard to find.

Why standardise the way that we run trials?

One of the goals of this book is to help people to run a more 'standard' form of randomised trial. This does not mean that all trials should be the same – they will need to be adapted to the local context as well as the nature of the intervention. Instead, standardisation means that where possible, the same *principles* of design should be adopted across studies. But why bother trying to do this at all?

The first and most important reason is one of practicality and cost. The more bespoke, or artisanal, a trial design is, the longer it will take to develop, and the more that will cost. We have an obligation to minimise spending on evaluations, conditional on them being conducted robustly, so standardisation is a good thing.

Next, one of the main risks to the validity of analysis of trials is that researchers and evaluators have to make a number of decisions, each of which can have a modest impact on the estimated effect size. This means that in many cases, you might, depending on your approach, get a different result. This is legitimate, and trials should vary to a certain extent. However, if you can choose your analysis based on the results it gives you – something which often termed the 'garden of forking paths'– this introduces bias in our estimates, and can make us think that things are more effective than they are.

Finally, comparability is useful for when we're making decisions. To the extent that different approaches to trial design and analysis give slightly different kinds of answers, then standardising the way we do things increases the extent to which we can make comparisons between different trials and different interventions, meaning that we can make better cases for investing in one thing over another. Larger-scale studies, called mega studies, which seek to test more interventions against each other, can help us establish direct comparisons between interventions (see chapter 14), but these are always going to remain a rarity.

What's next?

As we'll see over the coming chapters, there are a range of ways of running a trial. Over the remainder of this book, we're going to cover these, and try to provide you with the tools you need to run them. The book is divided into three parts:

Part One – which covers the main types of trial you might consider running, ranging from the simplest – individually randomised trials- through to the more complex, like stepped wedge trials, which exploit randomly ordered roll-outs of interventions to identify treatment effects.

Part Two – which covers the tools of the trade – from protocols, to power calculations, how to think about missing data, and considerations about ethics and consent. This section also goes into some novel approaches to trials, including winner-stays-on approaches and megastudies.

Part Three – which includes useful parameters that can be helpful in designing studies, such as intra-cluster-correlation rates and correlation rates for a range of outcomes and levels of potential randomisation in housing and homelessness and higher education.

Part 1
An introduction to randomised controlled trials

The first part of this book is focused on the main forms of randomised controlled trial that you might want to make use of, either as an evaluator/researcher, or as a commissioner. As such, each chapter considers a different form of RCT, starting at individually randomised trials (Chapter 1) and progressing through cluster randomised trials (Chapter 2), cross-over designs (Chapter 3), and finally to stepped wedges (Chapter 4).

Each chapter provides an illustration that type of trial and its key design features. Each type of trial faces different trade-offs. For example, individually randomised trials are the most statistically efficient and simplest to implement and explain. However, they also have the highest risk of spillovers.

By contrast, stepped wedge trials, in which the order of treatment is randomised, rather than participants being allocated to distinct treatment and control groups, is methodologically more complicated, requiring evaluators to deal with the potential confound of time; but it can also have ethical and practical advantages, particularly where a policy or intervention is going to be rolled out anyway.

In Part 1, we have attempted to provide intuitive explanations for the types of trial we've considering. More than the subsequent sections, these chapters are aimed at a relatively lay audience, who would like to find out 'what works', but don't know how, or which method might be most useful for their circumstance or policy.

Chapter 1. Introduction to individually randomised controlled trials

The most standard and basic form of trial

Randomised control trials (RCTs) are sometimes considered the best and most straightforward means of evaluating the impact of a programme or intervention. Among RCTs, individually randomised controlled trials (IRCTs) are the most basic and easy-to-understand type of trial – as well as the most reliable at identifying impacts, all else being equal. 'Individually' simply refers to the level of randomisation – in this case, at the individual participant level, for example individual students, job applicants, etc. This is in contrast to cluster-level randomisation (eg classrooms, schools) or randomisation by sites or regions, which will be covered in subsequent chapters.

Though conceptually speaking, IRCTs are considered the most basic trial, their implementation can be anything but basic. IRCTs must meet a high standard of theoretical criteria and implementation fidelity in order to reliably estimate effects, and as such, it is not always practical (or even possible) to implement them.

Simply put, IRCTs are most appropriate in cases where the treatment in question can be randomly allocated to individuals, spillover effects are unlikely, and an adequate sample size – or statistical power – can be achieved to detect outcome differences. Medical drug and treatment trials lend themselves well to an IRCT set-up, whereas in social policy domains like education, cluster randomisation is more common (because of the nature of interventions that often must be delivered to a whole class or even a whole school or because of concerns about contamination and spillovers). That said, understanding IRCTs is a key step to understanding other types of randomisation trials, so this chapter will cover the core concepts of IRCTs which underpin any causal inference claims, examples of IRCTs and how to build an IRCT protocol.

IRCTs: The basic procedure, why they are compelling, and important elements

As in any RCT, the goal of an IRCT is to estimate the effect of a treatment by comparing outcomes of a treated group against the outcomes of a control group, which represent the counterfactual case (ie what would have happened in the treatment group, absent treatment). The simplest version of this is a two-arm trial – one treatment arm and one control arm – but a trial can have three or more arms to test different treatment types, although this remains rare in most social policy domains.

The key element to any trial, regardless of how many arms it has, is the randomisation procedure, which in principle should ensure that every participant has the same probability of ending up in a given arm as any other participant. This means there is no correlation between a participant's characteristics – such as their age, gender, education, etc – and the treatment group they end up in.

The randomisation of treatment assignment performs a few important functions in your trial. First, it can allow the efficient and fair allocation of scarce resources (World Bank, 2016). In the case where demand for a particular treatment exceeds availability – for example, demand for tuition vouchers – a transparent process of randomisation ensures that each participant is just as likely as any another to receive treatment. Second, when done effectively, randomised treatment creates a higher probability of treatment and control groups being statistically identical, which is to say that each group is essentially the same across observed (eg age, education) and unobserved (eg motivation, preferences) characteristics that may impact the relevant outcomes. Third, and finally, transparent randomisation can enhance the credibility of researchers, both among trial participants and among policymakers and the scientific community assessing the findings. Participants can feel assured that no favouritism or corruption is at play, while policymakers and other researchers can feel secure in extrapolating findings to other contexts (this relates to a study's validity, discussed later in this chapter).

Elements of IRCT design

Here we will outline three important elements of IRCT design: randomisation protocol, statistical power, and balance tables. All of these should be documented as part of the pre-registration of a study and summarised in a write-up of findings.

Randomisation of treatment

Any randomisation protocol will likely be highly specific to the treatment and population of interest, but in general, the following two questions need to be answered as part of the randomisation design:

- **What is our population of interest, and is it possible to sample from it?** Are there specific characteristics we would need to satisfy when drawing our sample, and can these be verified at the data collection stage? In practical terms, drawing a sample from the population will often mean recruiting participants to join the trial. The vast majority of trials make use of either convenience sampling (recruiting those who are easiest to recruit), or targeted sampling (trying to recruit a group that is thought to particularly benefit from the intervention). Random sampling is rare, and not a requirement for a randomised trial (which depends on random allocation).

- **Can the treatment be randomly allocated as intended?** Is there any threat of spillover treatment into controls (for example, people in the treatment group sharing some element of the treatment with the control group – like homework resources being shared between parents of children in the same class), or non-random selection into the treatment group?

For example, let's say we are testing the impact of a training treatment on employment prospects for unemployed young people using job centre services. Our population of interest is a restricted group that requires certain defining characteristics to be met: young, unemployed, and visiting job centres. Thus we already know that our trial sample (ie study participants who will be sorted into treated and control groups) will need to draw from this population. Then, we need to ensure that our training treatment can be randomly allocated. Some questions surrounding this might include: Does training occur at specific times, possibly excluding some participants with restricted schedules? Does training require at-home internet access? Is the training offered in different languages? All of these could contribute to non-

random selection into the treatment group – and could therefore suggest exclusion criteria for the trial prior to randomisation taking place. Another related question is, does receiving treatment at all impact the potential outcomes of the control group, also known as spillover? For example, do study participants interact with one another, potentially allowing knowledge-sharing between the treated and control participants?

It may be that answers to these questions mean that an IRCT is not practical for assessing the training treatment; perhaps a quasi-experimental design, or cluster randomisation is better. But if these questions can be satisfactorily answered, there are a few strategies we can use to achieve randomisation.

A common method is through lottery: every participant is assigned a number from a range using a random number generator, then a number within the range is chosen, and everyone below that number receives treatment while everyone above is a control. Alternatively we can allocate using a coin toss (or, the computer simulated equivalent). In the end, the goal is to ensure that every participant has the same probability of being part of the treatment group as any other participant, which will give us the best chance of achieving statistically identical treatment and control groups.

External and internal validity

There are two types of validity: external and internal. For our trial to be externally valid, our sample needs to adequately represent the population of interest (ie our first question above), which means our trial's findings can be extrapolated to the wider population, otherwise known as 'generalisable'. Internal validity relates to our second question, on randomisation: does the control group accurately estimate the counterfactual outcome, or what would have happened with the treatment group if they never received treatment? High internal validity means that we can more safely assume that any difference in outcomes between the treated and control groups is attributable to the treatment.

As mentioned, the randomisation protocol lends transparency and credibility to participants and the scientific community, thus it is essential to scrutinise and document it thoroughly, with an eye toward supporting the claims of external and internal validity.

Trials are often argued to have high internal validity because they do not hinge on as many assumptions as other approaches to identifying causal impacts – but lower external validity than some other approaches because of the conditions that need to be met for randomisation to be successful (such as a self-selecting group of trial participants). The extent to which self-selection makes a difference that matters for extrapolating effects varies depending on the context. In a medical context, self-selecting participants generally share the same essential biology as non-participants, and so extrapolation might often be straightforward. In social policy, where motivation to participate might be highly correlated with the treatment effect, we might want to be more cautious.

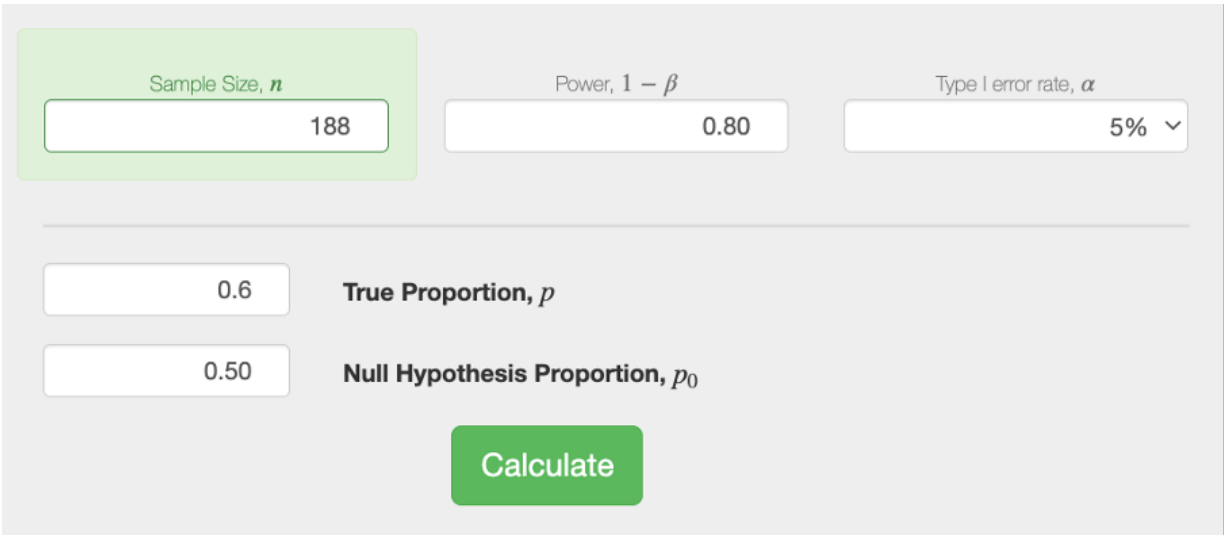
Statistical power

In concert with posing questions about our population of interest and randomisation protocol, an important early question to answer is: How large should the sample size be? This question relates to the concept of statistical power, which can be thought of as the probability of detecting an effect of treatment, if one exists. 'Power analysis' refers to the process of estimating the statistical power of a given study, and takes into account five things:

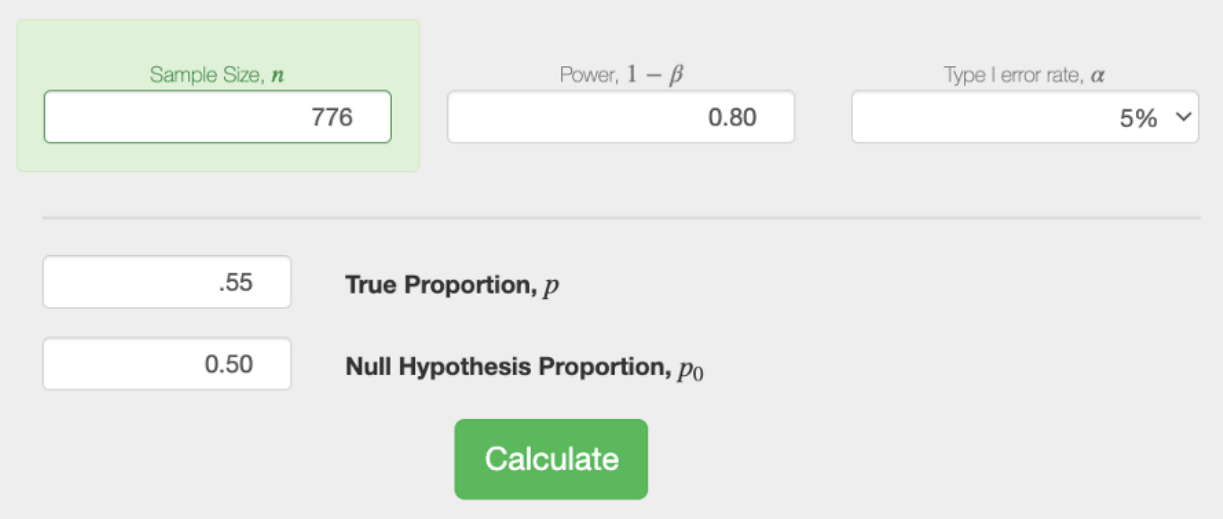
1. The effect size of interest.
2. The variation of X (ie the range of treatment; a two-arm trial would have a simple binary variation of 1 or 0).
3. The variation of Y (ie the standard deviation of measured data for our outcome of interest that is unexplained by the X treatment).
4. Statistical power (ie the probability of correctly rejecting the null hypothesis if an effect exists; 80 per cent is often conventionally used as the level of power, and the significance threshold is set at 95 per cent).
5. Statistical significance, denoted as alpha, is the probability of a type-one error (a false positive) assumed by the test, while power is one minus the probability of a type two error (a false negative), which is denoted as beta.
6. The sample size.

You will need any four items from the list in order to estimate the remaining fifth item. Power analyses done at the outset of a trial are usually used to project how large of a sample the trial needs (ie #5), while statistical power (#4) can be calculated after the study is completed – although this only useful under quite narrow circumstances. Since we are considering the design of IRCTs, we will focus on the former here.

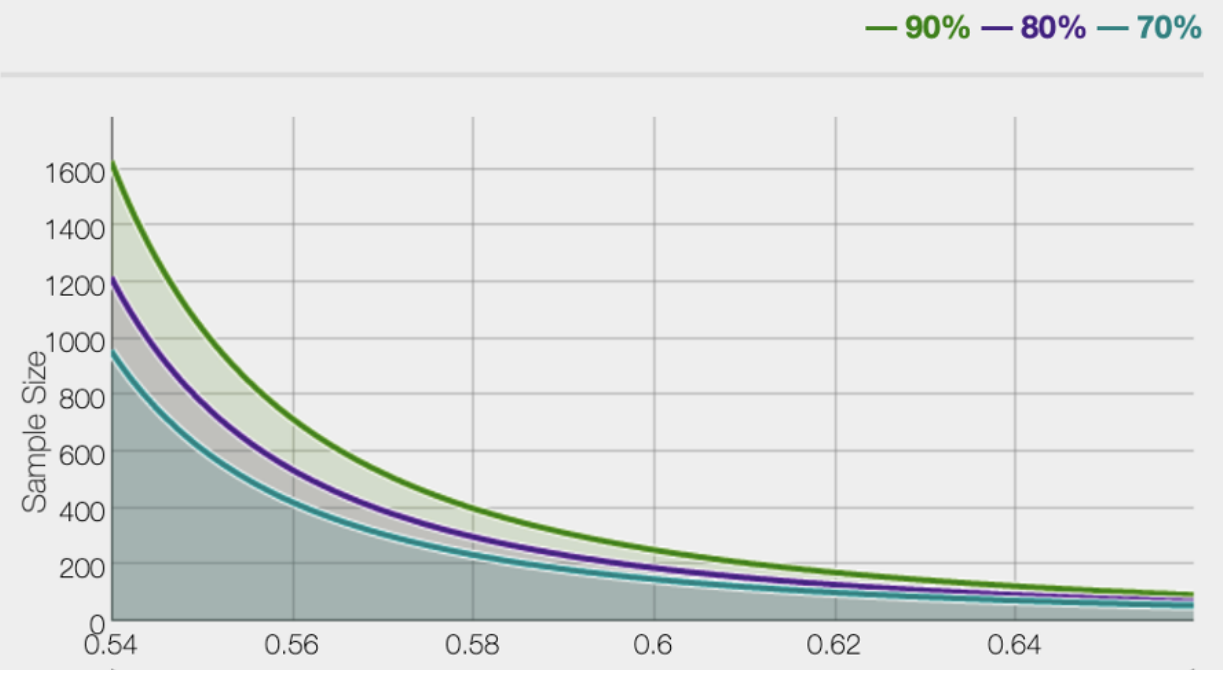
While power analysis can certainly be done by hand, or using statistical software such as R and Stata that are commonly available to evaluators, researchers, and government analysts, we can also take advantage of the many free, online calculators such as this one (HyLown Consulting, 2022). If our outcome of interest is a proportion, we can use this calculator.



After inputting what we think our 'true' proportion would be for treated participants (60 per cent, or 0.6 as a proportion) and the proportion of uptake for non-treated participants (50 per cent), we find that we would need a sample size of 188 participants to detect a 10-percentage point (pp) effect size with 80 per cent probability. What happens if the effect size is actuallysmaller than 10 pp? Let's try a five pp increase:



The required sample size increases substantially to 776. We can start to see a relationship emerging here: **as the anticipated effect size goes down, the required sample size goes up**. This makes intuitive sense, as we would expect that small differences in outcomes are easier to miss among all the observational noise than large differences in outcomes. The graph provided by the power analysis calculator helps to visualise this inverse relationship between effect size and sample size:



Note that this power analysis assumes a 50/50 split between treated and controlled participants. If your treated/control ratio is not an even split, you can take your ratio into account with this calculator. As one might expect, the more disproportionate our ratio is, the larger our sample will need to be. We should note as well that there are diminishing returns to a given increase in sample size in terms of the reduction in the effect size that is detectable.

It is difficult to give a robust estimate of the sorts of effect sizes we should "expect" to see in randomised trials, in part because in most policy domains the field is still in its infancy. As a general rule of thumb however, having the statistical power to detect effects of 0.2 standard deviations (also known as Cohen's d), is sensible – although if you have a particularly costly intervention, you may want to set this higher, and lower for a particularly low cost, light touch intervention. This book will be updated periodically to give estimates of effect sizes across domains, as more trial results become available.

Balance checks

Once we have completed our power analysis and implemented a randomisation of treatment protocol, we can strengthen our case that we have comparable groups to estimate our effect sizes through balance checks. A straightforward way to do this is to test (for example with a t-test) whether being in the treatment group is correlated with covariates: if age, gender, income, etc, correlate with selection into treatment, then we know that our groups are not balanced – that is, if the two groups are statistically different from each-other. If a statistically significant association is found, or one which is insignificant but nonetheless large in qualitative terms, we have a decision to make. We can either re-randomise, running our randomisation process again to achieve a result with better balance on our observable covariates, or we could accept the randomisation as is and control for these covariates in our final analysis to 'partial out' the effects of this variable. In chapter 5, we discuss the limitations of statistical significance as a method. Which approach to take is subject to disagreement, and for our purposes we recommend re-randomising, but noting that this has taken place in any reporting.

Estimating the effect

After writing our randomisation protocol and running our trial, we can work on estimating our effect. In principle, this is the simplest step. If our sample was sufficiently large, treatment and control groups were balanced, there was minimal non-compliance with treatment assignment within the sample, minimal missingness in the data collected, and no spillover effects, then we can simply take the average outcomes for the treated and control groups, then find the difference between them. If there is a difference, we then have evidence that the treatment had an effect.

That said, typically we want to include other variables that we believe may co-vary with our outcome and treatment variables, and we are also interested in the degree of uncertainty of our estimate. To do this, we can use statistical software such as Stata or R to run a regression model, which allows us to incorporate a multivariate approach and will give us a standard error and measure of statistical significance for our estimated treatment effect. Statistical significance simply refers to how confident we can be that our treatment estimate occurred because of the treatment, rather than by chance; typically we look for 95 per cent confidence, or a p-value of .05 (meaning a five per cent or less chance that the estimated difference is due to random chance). Different UK what works centres have different statistical analysis guidance that might be useful in specifying your own analytical approach, but some straightforward rules of thumb include:

- Making use of a form of regression analysis that is most appropriate for your outcomes – for example:
  - » Using logistic regression with binary outcomes.
  - » Using ordinary least squares regressions with continuous variables.
  - » Using Poisson regressions with 'count' variables.
- Controlling for stratification variables in your regressions.
- Controlling for variables that were imbalanced at the point of randomisation.
- Controlling for variables that are likely to strongly predict the outcome measure.

Examples of IRCTs

To better understand what IRCTs look like in the wild, let's cover a couple examples.

ParentChild+ trial

This Education Endowment Foundation-funded trial set out to assess the impact of a home visitation intervention programme known as ParentChild+. This was a two-arm trial (ie one treatment group, one control group), randomised at the individual family level, with the treatment consisting of a series of home visits aimed at increasing parent-child interaction. The population of interest was low-income families with children two to three years in age, and the study sample consisted of 320 families, evenly split between treatment control groups using a random number generator lottery system. The outcomes measured included the child's vocabulary acquisition (British Picture Vocabulary Scale) as well as other school-readiness indicators such as behaviour and fine motor skill development. The programme and trial were modelled after a trial conducted in the United States which had shown strong evidence of positive gains in child development; however, in the case of this trial, evaluators actually estimated a negative effect on relevant outcome measures of around two months' developmental delay compared with the control children.

Social prescription trial

This Cabinet Office-funded trial evaluated the effects of a light-touch, non-clinical intervention to increase people's well-being and social connection. The treatment was a type of 'social prescribing', whereby participants were prompted to increase their physical activity or social interactions with friends and family, with an aim toward improving wellbeing and social connection. 341 participants were recruited through Prolific and randomly assigned to one of four groups: three treatment variations, and one control group. The first treatment was a prompt to undertake physical activity, the second prompted social interaction, and the third similarly prompted social interaction but with a financial incentive of £10. Additionally, participants in the three treatment arms were asked to answer questions about how and when they were planning to engage in the proposed activities. Social connectedness and a scale of positive and negative experiences were measured at baseline as well as a week later, post-treatment. Evidence suggested statistically significant increases in the positivity of experience among participants in the physical activity and paid social connection arm.

Beyond the IRCT basics

Concealment and blinding

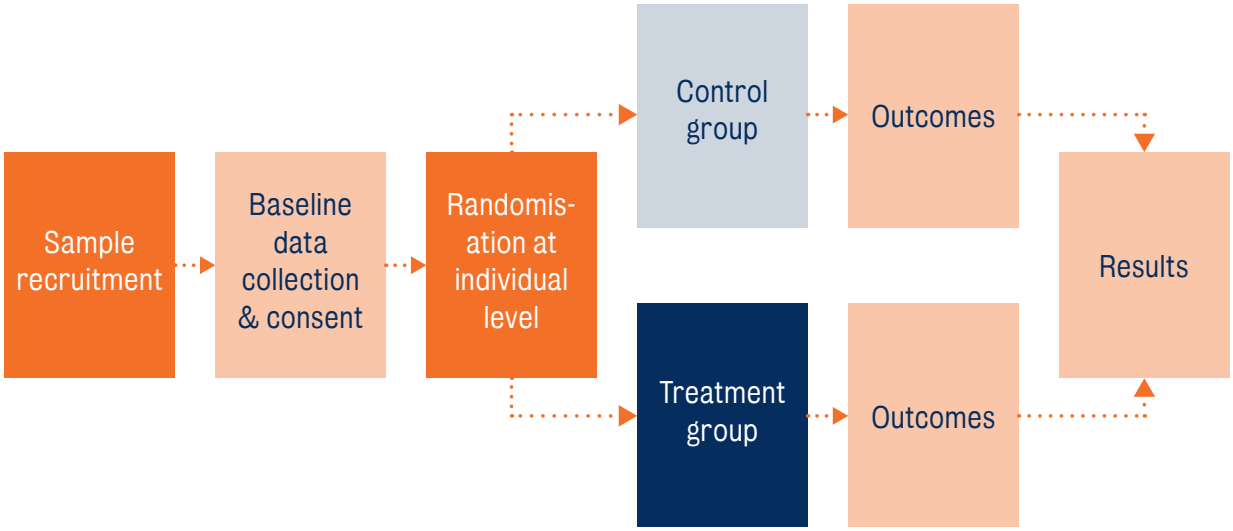
To prevent selection bias, it is essential that neither participants, researchers, nor implementation partners are aware of treatment assignments in advance. This process, known as allocation concealment, can be effectively achieved through strategies such as pre-sealed envelopes, which entails placing treatment assignments in sealed envelopes that are only opened after participants are enrolled.

Following randomisation, blinding can be employed to conceal the treatment assignment, ensuring that outcomes are not influenced by placebo effects or researcher bias. Studies can be single-blinded, in which the participants do not know which group they belong to (eg treatment or control group), or double-blinded, in which neither the participants nor the researchers know the group assignments.

Parallel designs

What we have described so far are IRCTs at their most basic and straightforward: two+ arm trials with a randomisation protocol that allocates treatment without conditions beyond what is stated in the population of interest. These types of IRCTs can be said to have 'parallel designs' where participants' assignment into treatment and control groups are consistent throughout the trial and results between the groups are compared at the end of the trial, ie in parallel with one another.

Figure 1 Parallel designs



Though they are straightforward to understand, there are some risks associated with parallel design. First, there are attrition risks, which is to say, participants who start out in the sample drop out at some point before the final outcomes are measured. This is a risk to all trials, but might be mitigated using a waitlist design, which we describe in the next section. This is not necessarily a threat to the trial validity; a certain amount of attrition should be expected, particularly with large samples and trials that take place over a long period of time. One risk comes from differential attrition, where members of either the treatment group or control group tend to attrit at higher rates than the other, or if participants with certain characteristics tend to drop out at higher rates than others. This can lead to attrition bias in your final effect estimates, threatening the internal validity of the trial and also possibly the external validity with respect to certain sub-populations. Attrition bias can be detected post-hoc through a missing at random (MAR) test, which essentially detects whether missingness is correlated with different observed characteristics. But wherever possible, it is far preferable to try to avoid differential attrition at the outset. One common way to do this is to make use of a waitlist design. Another threat to validity, even in the case of balanced attrition, is small sample sizes. If we lose too many participants, we also lose statistical power to detect reasonable effect sizes. Considering this, we should aim to recruit enough participants for a trial that means we will still have sufficient power, and randomisation is likely to be successful, once attrition has been accounted for.

Waitlist designs

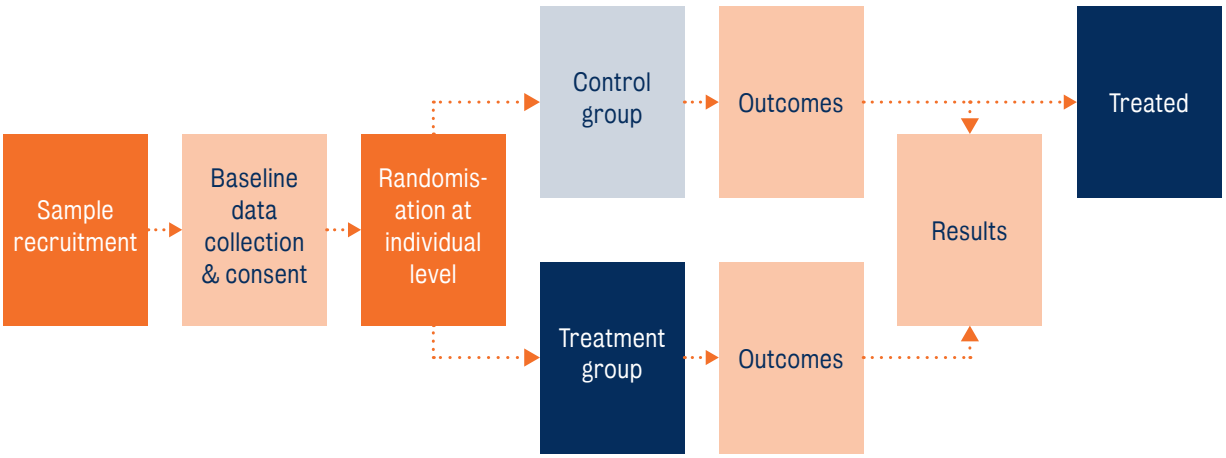
One possibility for mitigating differential attrition is through a waitlist design, wherein all trial participants will have an opportunity to take part in the treatment, but the control group in this case receives the treatment after the endline data collection. This can help address differential attrition among the control group as participants have more of an incentive to remain in the trial (provided of course that the treatment is considered desirable). For example, let's say we are interested in evaluating the effects of conditional cash transfers (CCTs) on school enrolment for girls in a low-income country context. During the participant recruitment stage, it would be much easier to encourage families to consent to join the trial and to remain engaged in the trial if it was clear up front that all qualifying participants will receive CCT payments eventually. In addition, waitlist trials might aid recruitment to the

trial overall, because participants will be guaranteed to receive the intervention – with some receiving it sooner than others – rather than only having, for example, a 50 per cent chance of receiving it.

Waitlist designs are also helpful in cases where the treatment needs to be implemented across an entire population, but the capacity for allocating treatment at a specific point in time is limited. For example, a ministry of education may need to roll out a new training programme with all teachers across the country. This potentially means training thousands of teachers, but given the capacity of the ministry and the costs associated, running a waitlist design IRCT can be beneficial for a couple of reasons: first, it is practical, respecting the training capacity of the ministry. Second, by randomly assigning the training treatment to a smaller group of teachers first, the ministry can estimate the impact of the training while also conducting a qualitative assessment of the training process, which can be used to improve implementation for future training cohorts of teachers. A more gradual rollout approach might give rise to a stepped-wedge trial, of which the waitlist trial is a special case. Stepped-wedge trials will be covered in more detail in chapter 4.

Waitlist designs are not without their drawbacks, however. Since all trial participants will eventually be eligible for treatment, this means we lose the ability to do long-term follow-ups to compare treated and control outcomes. If we suspect that our treatment will have delayed effects (as is commonly the case with early education interventions, for instance), waitlist designs will not be able to generate those insights. Another consideration with waitlist designs is whether the treatment will be appropriate for the control group at the end of the trial. If, say, our treatment is designed to be received at a certain developmental stage (as was the case with the ParentChild+ example discussed earlier), it may not make sense or be cost-effective to offer the treatment to the control group. Perhaps most importantly, a waitlist design is also expensive – because the funder will need to pay twice as much for the intervention to be delivered to everyone.

Figure 2 Waitlist designs

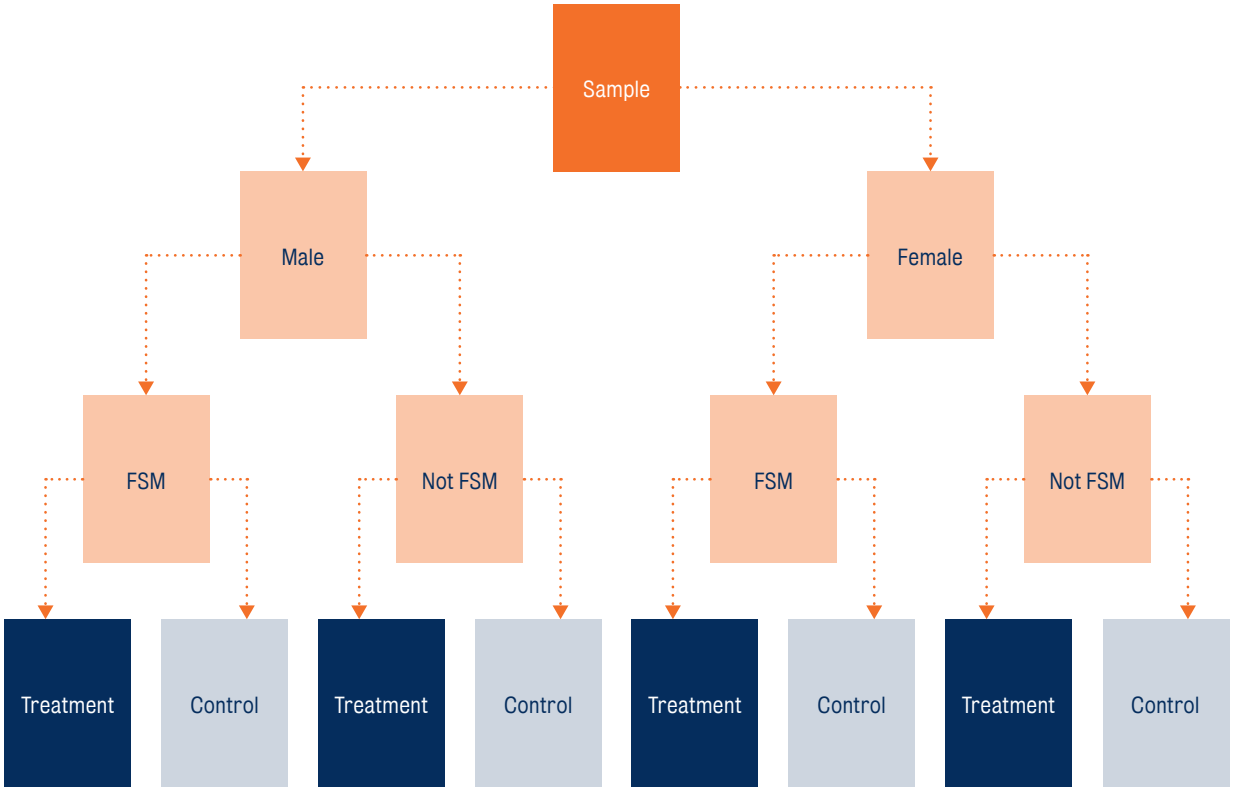


Stratification

There may be cases where we need to ensure individuals with certain characteristics are included in a trial, either because we believe the characteristic will be highly predictive of outcomes or because we judge that too little is known about certain subgroup effects. While randomised sampling should in principle represent characteristics in proportion with the population of interest, and randomised treatment assignment should see an equal distribution of characteristics between treatment and control, the realities of participant recruitment do not always support these outcomes. In these cases, we can implement stratification as part of the randomisation protocol.

Stratification is straightforward to understand: before randomising treatment, we split the sample by the characteristic of interest (say, gender or age grouping) and then we proceed with randomising the treatment assignment within the subgroups. This helps to ensure that our treatment allocation is balanced on these variables, rather than leaving it to chance. This extra step before randomising increases statistical power by ensuring greater balance, particularly in larger trials. In smaller studies, specifically those with fewer than 100 participants, stratification might reduce power.

Figure 3 Example stratification of pupils by sex and free school meals (FSM)



Oversampling

If one of our hypotheses explicitly relates to subgroup-specific effects (ie we want to test whether membership in a subgroup like women or minority children influences the effect of the treatment), then we will want to make sure our trial is sufficiently powered to capture this. This means ensuring a sufficient number of participants from the subgroup of interest are recruited for the trial. This can be achieved through oversampling, or recruiting more participants from the subgroup of interest. More participants from a subgroup of interest enable us to estimate treatment effects within that subgroup with more certainty. We may do this particularly if we are concerned about the equity effects of an intervention that is known to be especially likely to produce negative outcomes. Early trials funded by the Education Endowment Foundation imposed recruitment targets, particularly for schools that ranked high on the Income Deprivation Affecting Children Index (IDACI), and hence deliberately oversampled schools where children were from low-income families.

Note that if we oversample, this effectively means that our sample may no longer represent the population, which potentially threatens the external validity of our trial. This imbalance can be addressed by applying weighting, but most importantly, any oversampling methods should be documented thoroughly in both the trial protocol and the write-up of results.

When deciding how much to oversample by, there are two things to consider. First, some groups may experience differential attrition from the study than others – for example, people with lower baseline scores might be more likely to drop out of the study than their peers. Where this is the case, you might wish to oversample to the extent that attrition differs between the two groups, to maintain a sample for analysis that represents both groups. Second, if you are interested in looking at subgroup effects (that is, the effects of the intervention for the group being oversampled), you might wish to conduct separate power calculations for this subgroup to ensure you can detect reasonable effect sizes. In practice, detecting differences between effects experienced by different subgroups is likely to require very large samples, perhaps four times that for detecting the main effect, and so may be impractical.

Ethics and critiques of RCTs

Many introductions to RCTs begin by referring to RCTs as the 'gold standard' of programme evaluation. This is for good reason, as RCTs are highly effective at estimating treatment effects when they are implemented properly. However, RCTs are not without hazards, and due in part to their popular implementation across fields, many organisations have set up internal review board (IRB) approval processes that vet RCT proposals on ethical considerations. In a university context in the UK, this is likely to involve a research ethics committee, but for other evaluators, boards may take different names and forms. Every IRB process will be different, depending on the population in question and the jurisdictional context, but in general, they will investigate trials in terms of their equipoise, informed consent, data protection and minimisation (the need to avoid collecting too much data from participants), and accountability (ensuring researchers can be held accountable for decisions).

Equipoise refers to a genuine uncertainty about the benefits of a certain treatment, which necessitates the testing of treatment through randomisation (IPPR, 2021)(IPPR, 2021). The simplicity of this definition belies the complex moral questions around potential harms from withholding treatment to study participants, which is often one of the biggest concerns stakeholders might have around RCTs. This concern can in part be ameliorated through a waitlist design as discussed, but the core question IRBs are interested in is whether the trial is truly necessary in order to advance understanding, and that taking part in either a treatment or control group is reasonably unharmful for participants. However, when considering equipoise, it is important to distinguish between strength of feeling and the actual evidence. The overwhelming majority of interventions that are to be tested have supporters who believe passionately that they work. Nonetheless, the bulk of evidence from the what works network thus far shows that many of these interventions are ineffective.

Informed consent is another area of deep complexity, but in general, participants should be briefed on the purpose and methodology of the trial, how collected data will be used, and whom they can contact with any concerns or questions. Furthermore, participants should be able to revoke consent at any time and should be aware of the processes to do so. The standards of consent for medical treatments may often not be appropriate for a social policy trial, for example one where the alternative to a trial is not that nobody receives the intervention, but rather that it is rolled out without an evaluation. This argument applies less for trials where individual randomisation is possible, as individual recruitment suggests that individual consent might also be feasible. We cover some of the nuances around information consent, as compared to informed assent, in chapter 9.

Data protection and minimisation mean that data gathered from participants, particularly personally identifiable information (PII), needs to be adequately safeguarded and the principle of data minimisation should be respected, which is to say that any data collected should be strictly necessary for the study and should be properly deleted when no longer needed.

Accountability means that evaluators should make themselves appropriately accountable to participants, which can include any and all of the following: providing transparent and ongoing communication throughout the trial, being responsive to questions and concerns from participants, and sharing preliminary findings if appropriate.

Critiques of RCTs

In the push for more quantitative, evidence-based research in the social sciences, RCTs have grown more popular and have found applications across diverse sectors. This has led some to comment that a 'hierarchy of methodologies' is being adopted, with RCTs held in higher esteem than quasi-experimental designs (QED), which are in turn preferred over more observational methods (IPPR, 2021). This potentially has implications for which studies receive funding, and which are prioritised for journal publication, etc. It can be tempting to say that since RCTs provide excellent estimates of the counterfactual, they should be used as much as possible, to the exclusion of other methods. However, there are numerous, unignorable reasons why an RCT may not be an appropriate evaluation method.

One such reason is when the treatment (or withholding thereof) may pose significant harm to participants. A straightforward example of this is evaluating the impact of second-hand smoke on children. The treatment of second-hand smoke is neither practical nor ethical to randomise. In this case, a quasi-experimental approach such as matching would be a better option.

Another situation where we might reconsider doing an RCT is in the case of already highly studied populations. Due in part to the growth of RCT usage around the world, a problematic phenomenon known as 'participant fatigue' is becoming more common. In international development contexts, both Kenya and India have emerged as RCT hotspots, leading to some unintended consequences (IPPR, 2021). Consider the experience of living in a small town or village with an administration friendly to international NGOs; there may be research teams coming and going, always recruiting for trials. Participants learn over time how to strategically interact with these studies, which can lead to misreporting and bias. Furthermore this isn't exactly a dignified experience, and feelings of alienation between researchers and study participants can be common. These experiences are not confined to international development, of course – minorities and other traditionally marginalised people residing in rich countries report similar feelings of being overstudied, and often are left out of conversations around appropriate policy interventions and research approaches (Chicago Beyond, 2019).

Conclusion

IRCTs are a powerful tool and remain one of the best ways to estimate treatment effects. However, IRCTs must also meet rigorous requirements in order to claim high external and internal validity and produce reliable estimates. These requirements include, but are not limited to, the ability to randomise treatment, a sufficiently powered sample, high-fidelity data collection, minimal participant attrition, and an ethical context conducive to randomised testing. Not all trial contexts will be practical for an IRCT. In subsequent chapters, we will cover various ways to mitigate barriers to RCT implementation, as well as alternatives when RCTs are simply not possible or desirable.

References

Chicago Beyond. (2019). 'Why am I always being researched? Guidebook.'
https://chicagobeyond.org/researchequity/

Education Endowment Foundation. (2021). Trial Evaluation Protocol: ParentChild+.
https://d2tic4wvo1iusb.cloudfront.net/documents pages/projects/ParentChild_EEF_trial_protocol_v1.3_15022021_FINAL.pdf?v=1680612491

HyLown Consulting LLC. (2022). Power and Sample Size.
http://powerandsamplesize.com/

Menon, S. (2021). 'RCTs for policy in India – Ethical Considerations, Methodological Concerns and Alternative Approaches'. Indian Public Policy Review. 2(3): 65-86.
https://ippr.in/index.php/ippr/article/view/48/28

World Bank (2016). Impact Evaluation in Practice.
https://www.worldbank.org/en/programs/sief-trust-fund/publication/impact-evaluation-in-practice

Chapter 2. Cluster randomised controlled trials

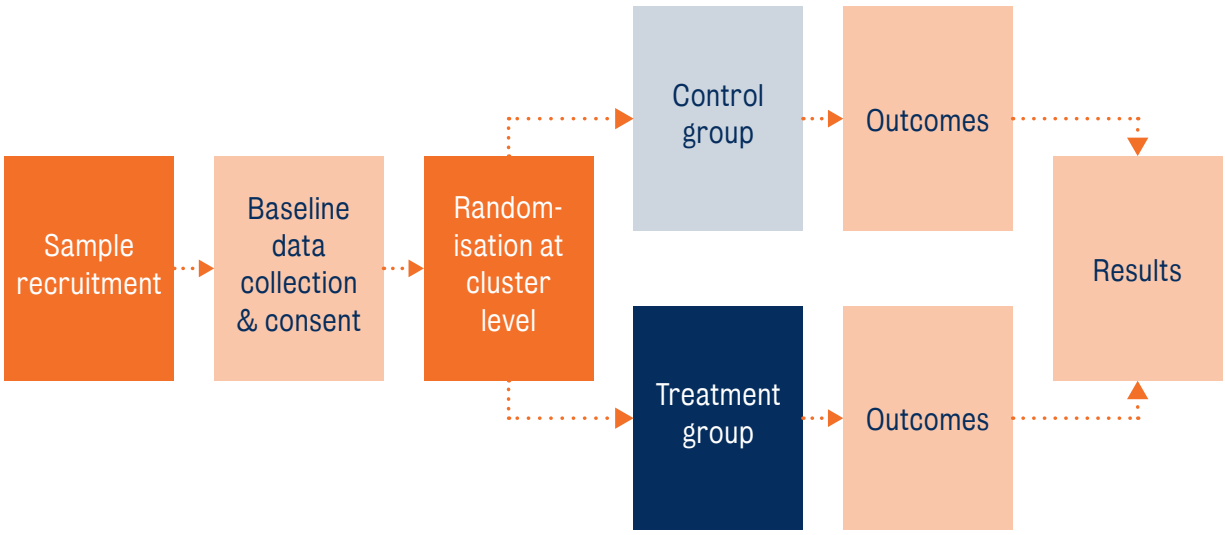
The basic structure, when to use them, and important elements

Cluster randomised controlled trials (CRCTs) are a more complex trial design used to evaluate the impact of a programme or intervention. Unlike IRCTs (discussed in the previous chapter), where individual participants are randomly assigned to treatment or control groups, the unit of randomisation in CRCTs is the group or 'cluster'.

CRCTs are typically employed when the nature of the intervention is administered to groups rather than to individuals (eg a public health campaign or classroom curriculum), when there's a risk of spillover or contamination, or when ethical and practical considerations make individual randomisation challenging. Examples of clusters include schools, regions, households, hospital wards, and so forth. Once clusters are identified, each cluster is then randomly assigned to either a treatment or control arm – the choice of the clustering unit depends on the nature of the intervention and how it is delivered. However, you want to randomise at the lowest possible level while reducing contamination – that is, the likelihood that people in the treatment group have some effect the outcomes of the control group, or people in the control group end up receiving some dosage of the treatment (Eldridge & Kerry, 2012; Hemming & Taljaard, 2023).

Once clusters are defined, the next step in CRCTs is to randomise the cluster to either the treatment or control group. As with IRCTs, randomisation can be performed using simple or stratified randomisation techniques. An example of a simple randomisation approach for a CRCT is shown below. In this case, the unit of randomisation is the cluster (eg hospitals), with entire hospitals either allocated to the treatment or control group.

Figure 1 Cluster randomisation



While in an IRCT individuals are randomly selected, and therefore have independent observations across outcomes, in a CRCT, there is often correlation among the characteristics and outcomes of individuals within the same cluster. This degree of correlation within a cluster is quantified using the intra-cluster correlation coefficient (ICC), which must be accounted for when calculating sample size in a CRCT. Since the assumption of independence of observations is violated, CRCTs generally require larger sample sizes than IRCTs to maintain the same statistical power (Eldrige et al., 2009).

Sample size and statistical power

In choosing sample size for a CRCT, you need to determine the total number of clusters required and the number of individuals per cluster. To do this, first use the strategies outlined in chapter 1 to calculate the entire sample size for an IRCT. Next, inflate the IRCT sample size to account for clustering by calculating the design effect using the formula below.

$$\text{Design Effect} = 1 + (M-1) \times ICC$$

Where 'M' is the size per cluster and 'ICC' is the degree of correlation between individuals within a cluster. The ICC is between 0 and 1 and can be estimated using relevant published literature or pilot data, if available (Chapter 11 provides estimated ICC rate tables for outcomes in homelessness and higher education). A higher ICC indicates stronger correlation between individuals in the same cluster.

Once you have the design effect, multiply the IRCT sample size by the design effect to get the total sample size required for the CRCT. Finally, to calculate the required number of clusters, simply divide the total required sample by the average cluster size.

One of the key considerations impacting the required sample size is the balance between the number of clusters and the size of each cluster. Increasing the number of clusters tends to have a greater impact on power than increasing the size of clusters, since each additional cluster contributes independent information to help detect differences between the treatment and control groups. Therefore, when planning a CRCT, it is often more efficient to aim for a larger number of smaller clusters rather than fewer larger clusters. The most simple CRCT design includes equal sized clusters, however, more advanced methods can be used if this assumption does not hold (Dron et al., 2021).

Estimating the effects

In CRCTs, effects can be measured using two approaches. The first method is cluster-level analysis, in which you aggregate cluster-level outcomes using summary statistics and then compare cluster means or proportions using t-test or regression models. While simple, this approach is limited by its inability to account for individual-level covariates or reveal individual-level treatment outcomes.

The second, more conventional approach for CRCTs is to estimate effects for individuals within clusters. Mixed-effect models such as general linear mixed models (GLMM) or generalised estimating equations (GEE) can be well-suited for accounting for observed and unobserved within-cluster correlations. Alternatively, if the CRCT incorporates repeated measures over time, autoregressive (AR(1)) modelling can be used. The choice of your statistical analysis will depend on the data you have available (eg time series, panel), the sample size, and the cluster sizes and number of clusters.

As with IRCTs, you can use individual-level covariates (eg age or sex) to improve power. Any estimation method must also account for the ICC, or risk type I error rates; therefore calculating cluster-robust standard errors is common practice.

Examples of CRCTs

Let's cover a few examples of appropriate uses of CRCTs:

Police in classrooms

This Youth Endowment Fund pilot trial, carried out by King's College London and Cardiff University, utilised an CRCT to examine the impact of a new PSHE curriculum co-taught by teachers and police officers on young people's offending behaviour, behavioural difficulties, and trust and confidence in police. Since the treatment (ie three PSHE lessons co-taught by teachers and police officers) was delivered to classes during their PSHE class, an IRCT design was not possible as there was no practical way to individually randomise which students in the class received the PSHE lesson, and which did not. The study instead randomised the treatment and control groups by year (ie all classes within a given year were either taught the new curriculum or continued with business as usual). The decision was made to randomise at the year level rather than the classroom level to reduce contamination, as it was hypothesized that students were more likely to speak to students in their year about the intervention than students in other year groups. Police administrative data and baseline and endline surveys were used to evaluate the impact of the intervention.

STOP Colon Cancer

The Strategies and Opportunities to STOP Colon Cancer in Priority Populations study (Coronado et al., 2014) used a CRCT design to test the impact of an electronic health record programme on colorectal cancer screening rates in 26 clinics across Oregon and Northern California. Randomisation was stratified by clinic organisation with an equal number of clinics being randomised into the treatment and control groups. The study used two treatment arms: in the first, clinics in the group identified patients due for colorectal cancer screening and automatically mailed them testing kits; in the second, clinics implemented an improvement process to optimise the programmes adoption. The control clinics continued with usual care. Outcomes included proportion of eligible patients completing a test within 12 months, as well as cost-effectiveness and return on investment of the programme.. A CRCT design approach was appropriate in this study as it would have been impractical and expensive to implement an automated record system for only certain individuals within a clinic. Depending on its success, the STOP Colon Cancer study has potential to offer a scalable and cost-effective model for increasing colorectal cancer screening rates.

Risks, threats to validity

While CRCTs are advantageous in evaluating interventions at the group level, it is important to consider the risk this design approach poses to bias, validity, and operational feasibility. In comparison with a standard IRCT, due to ICC, CRCTs can require much larger sample sizes to achieve the same statistical power, which often renders them expensive to implement and logistically challenging to manage.

CRCTs are also prone to greater selection bias than IRCTs as bias can be introduced at the individual or the cluster level. Since individuals within clusters are not randomly assigned to a treatment or control conditions, individuals within clusters may share characteristics that systematically differ from individuals in other clusters. There is also a higher chance of imbalance in baseline characteristics between study arms since randomization occurs at the cluster level, which can threaten internal validity. As with IRCTs, attrition bias remains an issue in CRCTs and can be compounded by variation in differential loss between clusters. Furthermore, CRCTs sometimes prompt generalisability concerns as they are conducted in specific settings (eg schools, hospitals, geographic regions) which may not translate to the general population.

While CRCTs must grapple with similar ethical challenges as IRCTs, there is an added layer of complexity with who should be providing consent, since treatment randomisation is assigned to groups. It can be particularly difficult to ensure that all participants are fully informed about the study's nature, risks, and potential benefits in CRCTs as individuals are sometimes unaware they are even part of a study (eg if a health programme is being trailled in their hospital). Moreover, obtaining individual consent can be logistically challenging and may risk introducing bias if those who consent differ significantly from those who do not between different clusters. Therefore, in CRCTs, informed consent is often only provided by gatekeepers of the cluster (eg principals, community leaders, managers) rather than individual participants. Researchers must navigate balancing the ethical requirement for informed consent alongside the feasibility of rolling out a CRCT (Eldridge & Kerry, 2012; Farrugia, 2021; Hayes & Moulton, 2009; Puffer et al., 2005).

Given these limitations, CRCTs should be avoided when individual randomisation is ethically, logistically, and statistically possible. However, if a trial more appropriately lends itself to a CRCT design, strategies can be used to address the aforementioned challenges, such as adjusting for baseline differences among study arms, using statistical techniques such as mixed models to account for ICC, and ensuring CRCTs have large enough sample sizes to detect effects.

Conclusion

CRCTs are a useful alternative to IRCTs when evaluating interventions that are administered at a group or community level. They are particularly valuable in public health, education, and other fields where interventions are naturally applied to groups and individual randomisation is not feasible. Many of the same design, data collection, randomisation, sampling, and ethical strategies from IRCTs can be applied to CRCTs. However, it is imperative that the correlation among individuals within a cluster (ie the ICC) is accounted for at all stages of the CRCT design and analysis.

References

Brown, A. W., Li, P., Bohan Brown, M. M., Kaiser, K. A., Keith, S. W., Oakes, J. M., & Allison, D. B. (2015). Best (but oft-forgotten) practices: designing, analyzing, and reporting cluster randomized controlled trials. The American journal of clinical nutrition, 102(2), 241–248.
https://doi.org/10.3945/ajcn.114.105072

Coronado, G. D., Vollmer, W. M., Petrik, A., Taplin, S. H., Burdick, T. E., Meenan, R. T., et al. (2014). Strategies and opportunities to STOP colon cancer in priority populations: Design of a cluster-randomized pragmatic trial. *Contemporary Clinical Trials*, 38(2), 344–349.
https://doi.org/10.1016/j.cct.2014.06.010

Dron, L., Taljaard, M., Cheung, Y. B., Grais, R., Ford, N., Thorlund, K., Jehan, F., Nakimuli-Mpungu, E., Xavier, D., Bhutta, Z. A., Park, J. J. H., & Mills, E. J. (2021). The role and challenges of cluster randomised trials for global health. *The Lancet Global Health*, 9(e701–e710).

Eldridge, S., & Kerry, S. (2012). *A practical guide to cluster randomised trials in health services research*. Wiley.

Eldridge, S. M., Ukoumunne, O. C., & Carlin, J. B. (2009). The intra cluster correlation coefficient in cluster randomized trials: a review of definitions. *International Statistical Review*, 77(3), 378-394.

Farrugia, P. (2021). Statistics in brief: The cluster randomized controlled trial–What is it and why is it relevant to research in surgery? *Clinical Orthopaedics and Related Research*, 479(8), 1852–1857.
https://doi.org/10.1097/CORR.0000000000001859

Hayes, R.J., & Moulton, L.H. (2009). *Cluster Randomised Trials* (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781584888178

Hemming, K., & Taljaard, M. (2023). Key considerations for designing, conducting and analysing a cluster randomized trial. *International Journal of Epidemiology*, 52(5), 1648–1658.
https://doi.org/10.1093/ije/dyad064

Puffer, S., Torgerson, D. J., & Watson, J. (2005). Cluster randomized controlled trials. *Journal of evaluation in clinical practice*, 11(5), 479–483.
https://doi.org/10.1111/j.1365-2753.2005.00568.x

Chapter 3. Crossover (within-subject) trials

Introduction

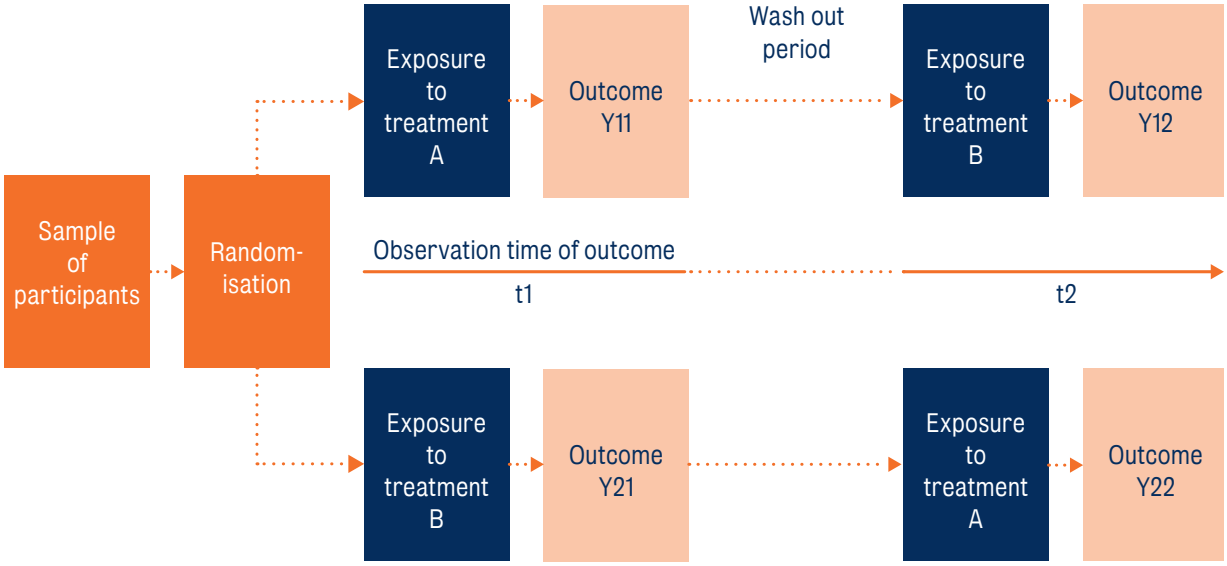
Crossover (or within-subject) trials are randomised trials in which each subject serves as their own control in the study, or where a cluster can act as its own control. Their design is particularly well-suited for conditions where the treatment effect is temporary while the outcome remains relatively stable over time (Sedgwick, 2014), or, where there are clear differences in cohorts. For example, we can think of a trial of a hot-spot policing trial. In this kind of trial, local areas are assigned to either have an extra, visible police presence, or not to. In a cross-over trial, this could be alternated, with some postcode areas receiving the extra policing one month, and then not the next, while other areas experience the opposite.

Basic design

In a crossover trial, participants are randomly assigned to different study arms, where each arm follows a sequence of two or more treatments administered sequentially. The simplest example of this is the AB/BA design. Assuming two treatment conditions, A and B, then participants are randomly allocated to be exposed to both conditions in a different sequential order. In the AB arm, one group of participants are exposed to condition A first, followed by condition B, while for the second group the treatment sequence order is reversed with exposure to the BA arm.

Crossover trials enable a comparison of a participant's outcome response to condition A based on their own response to condition B instead of comparing them to another participant exposed to condition B as in parallel designs. Usually, there is a gap in time between each exposure to each treatment condition called a 'wash-out period', which allows for any remaining carryover treatment effects to be fully eliminated prior to exposure of the next treatment condition.

Figure 1 Crossover design



By accounting for individual variability, crossover trials can be more efficient than similarly sized parallel randomised control trials, where each participant is randomly exposed to only one arm. Since each participant serves as their own control, this design provides greater statistical power for the same sample size, allowing treatment effects to be estimated with greater precision.

Uniformity and balance

Crossover trials can extend to more complex designs with multiple periods and arms, allowing researchers to achieve balance through careful planning. Achieving uniformity and balance in the design helps mitigate confounding factors such as sequence, period, and carryover effects.

- **Uniform designs:** A crossover design is uniform within sequences when each treatment appears an equal number of times in each sequence. This ensures sequence effects are unconfounded with respect to treatment differences. Similarly, when the design is uniform within periods, each treatment appears an equal number of times in each period, ensuring period effects are not confounded with treatment differences. For example, a uniform design might be on in which there are 30 schools in a trial, and two time periods, in each of which 15 schools receive treatment and 15 do not.

- **Balanced designs:** design is considered balanced for first-order carryover effects when each treatment is preceded and followed equally often by every other treatment. This helps eliminate confounding between carryover effects and treatment differences, provided the carryover effects are homogeneous between treatments. In the context of an individually randomised crossover design with three conditions (Control, Treatment 1 and Treatment 2), there will be as many people whose treatment sequence is Control ->Treatment 2 -> Treatment 1, as there are whose sequence is Treatment 2 -> treatment 1 -> Control – and the same for other conditions. Having a design that is balanced in this way helps ensure that any lingering effects of the intervention from the previous period are accounted for across the study design. In a two armed crossover trial, or one, as in schools, where there are different cohorts of participants in each time period (but a common set of clusters), this balance is achieved automatically.

- **Strongly balanced designs:** A strongly balanced design extends a balanced design by ensuring that each treatment precedes and follows every other treatment, including itself, an equal number of times. This is typically achieved by repeating the last period in a balanced design. Strong balance ensures that first-order carryover effects are accounted for, eliminating potential confounding with treatment differences.

Irrespective of the strength in balance, a wash-out period is usually advised, since the assumption that carryover effects are completely eliminated in a balanced design relies on specific conditions that are not always met in practice.

Latin squares

A Latin Square is a table, composed of the same number of columns and rows, in which each element exists exactly once in each row, and in each column. A popular example of a Latin Square is a completed sudoku puzzle – in which each of the numbers 1-9 appear exactly once in each row, and exactly once in each column.

We can think of a Latin square design when we are designing a crossover trial with more than two conditions.

As observed with the simpler AB/BA design, the Latin square can also provide a foundation for more complex k-period, k-treatment crossover designs – that is, where the number of time periods is the same as the number of treatments we're testing. This approach ensures a uniform design where each treatment occurs exactly once within each sequence and each period. For example, in a three-treatment (A,B,C) and three-period design, the Latin square sequences would be: ABC, BCA, CAB. In this square design, we can see that in each time period (position within each sequence of letters), each treatment exists, and that within each sequence of three letters, all three treatments exist.

In a Latin square design:

- Uniformity across **periods** ensures that period effects are removed.
- Uniformity across **sequences** ensures that sequence effects are removed.

This structure minimises confounding factors and allows for precise treatment comparisons in a simple and efficient design.

Williams design

The Williams design is an extension of the Latin square design that accounts for first-order carryover effects for when we have three or more conditions. It ensures that each treatment is preceded and followed equally often by every other treatment, leading to a balanced design. For the same three-treatment example, the sequences would expand to two Latin square designs: ABC, BCA, CAB, ACB, BAC, CBA. The design includes k(k−1) possible treatment pairs for all k treatments, where each pair has a corresponding 'opposite pair' that cancels out carryover effects – so, if there are three treatments in a study, there are 6 = (3 x (3-1)) combinations.

General rules for Williams designs:

- When the number of treatments k is even, a single Latin square can yield a balanced design.
- When k is odd, two Latin squares are required to achieve balance.
- A balanced Williams design ensures that each treatment is preceded by every other treatment equally often, cancelling out first-order carryover effects.

By accounting for carryover effects, the Williams design improves the accuracy of treatment effect estimates, especially when carryover effects might vary depending on treatment order.

Advice on analysis

Since crossover trials involve repeated measures, each participant has outcomes under different conditions. This paired data structure allows for within-subject comparisons, reducing variability. The most common statistical method for analysing crossover trials is ordinary least squares (OLS) or a mixed-effects model.

Two-treatment, two-period crossover

For the simplest (AB/BA) design, analysis often involves:

- Paired t-tests: If there is no significant period or carryover effect, a paired t-test can compare the means of the two treatments across participants.
- OLS with fixed effects and interactions: This can model both treatment and period effects and is suitable when assumptions of normality and equal variances hold.

Example

A within-subject crossover trial by Norman et al. (2018) examined the sustained effects of unhealthy food advertising on children's dietary intake. The study was conducted across four school holiday camps in Australia over a period of six days with a sample of 160 children aged seven to 12 years. Each camp included two gender- and age-balanced groups, which were randomised to either a single or multiple media condition and exposed to food and non-food advertisements via TV cartoons and/or online games. A randomised, crossover design was employed, with children experiencing both advertising conditions in counterbalanced order. Food consumption was then measured after snacks following the advertising exposure, and again at lunch later in the day. Children exposed to multiple-media (TV and online) food advertising ate more snacks compared to those exposed to non-food advertising, and the combined effect of online and TV advertising was stronger than TV advertising alone. Results showed children with baseline heavier weight status had higher food intake in both conditions.

Summary of design features

The impact of various design features can be summarised in the following points:

- When a crossover design is uniform within sequences, sequence effects are not confounded with treatment differences.
- When the design is uniform within periods, period effects are not confounded with treatment differences.
- In a balanced crossover design, carryover effects are not confounded with treatment differences, provided the carryover effects are homogeneous between treatments.
- In a strongly balanced design, carryover effects are fully accounted for in treatment differences.

Despite their efficiency, crossover trials require meticulous planning to address potential challenges which can compromise the validity of results. When carryover effects are suspected, a wash-out period is generally recommended even under a balanced design, in order to account for potential heterogeneity of treatment.

The analysis of crossover trials often leverages statistical methods like paired t-tests, repeated measures Analysis of Variance (ANOVA), or mixed-effects models, depending on the complexity of the design and the presence of period or carryover effects. When properly designed and analysed, crossover trials provide a robust framework for investigating interventions, maximising both resource efficiency and the reliability of findings.

Conclusion

Crossover trials are a powerful and efficient design for evaluating interventions, particularly when treatment effects are temporary, and outcomes remain stable over time. For example, this could include nudge trials that increase the salience of an option or that reinforce a particular behaviour like physical activity (Bondaronek et al., 2021); the implementation of a future policy in education, such as the effect of improved school meals (Widenhorn-Müller et al., 2008); or the provision of financial incentives such as small cash transfers amounts (Lippman et al., 2023). By allowing participants to serve as their own controls, crossover trials offer greater statistical power and precision with fewer participants compared to parallel randomised control trials, while simultaneously adhering to ethical considerations of fairness in the allocation of the intervention.

References

Bondaronek, P., Slee, A., Hamilton, F., & Murray, E. (2021). The public health potential of two popular apps to increase physical activity. *European Journal of Public Health*, 31(Supplement_3), ckab164.506. https://doi.org/10.1093/eurpub/ckab164.506

Lippman, S. A., Libby, M. K., Nakphong, M. K., Arons, A., Balanoff, M., Mocello, A. R., Arnold, E. A., Shade, S. B., Qurashi, F., Downing, A., Moore, A., Dow, W. H., & Lightfoot, M. A. (2023). A guaranteed income intervention to improve the health and financial well-being of low-income Black emerging adults: Study protocol for the Black Economic Equity Movement randomized controlled crossover trial. *Frontiers in Public Health*, 11, 1271194. https://doi.org/10.3389/fpubh.2023.1271194

Mills, E. J., Chan, A. W., Wu, P., et al. (2009). Design, analysis, and presentation of crossover trials. *Trials*, 10(1), 27. https://doi.org/10.1186/1745-6215-10-27

Norman, J., Kelly, B., McMahon, A.-T., Boyland, E., Baur, L. A., Chapman, K., King, L., Hughes, C., & Bauman, A. (2018). Sustained impact of energy-dense TV and online food advertising on children's dietary intake: A within-subject, randomised, crossover, counter-balanced trial. *International Journal of Behavioral Nutrition and Physical Activity*, 15(1), 37. https://doi.org/10.1186/s12966-018-0672-6

Sedgwick, Philip. (2014). What is a crossover trial?. *BMJ* (online). 348. g3191. 10.1136/bmj.g3191.

Widenhorn-Müller, K., Hille, K., Klenk, J., & Weiland, U. (2008). Influence of having breakfast on cognitive performance and mood in 13- to 20-year-old high school students: Results of a crossover trial. *Pediatrics*, 122(2), 279–284. https://doi.org/10.1542/peds.2007-0944

Chapter 4. Stepped-wedge randomised controlled trials

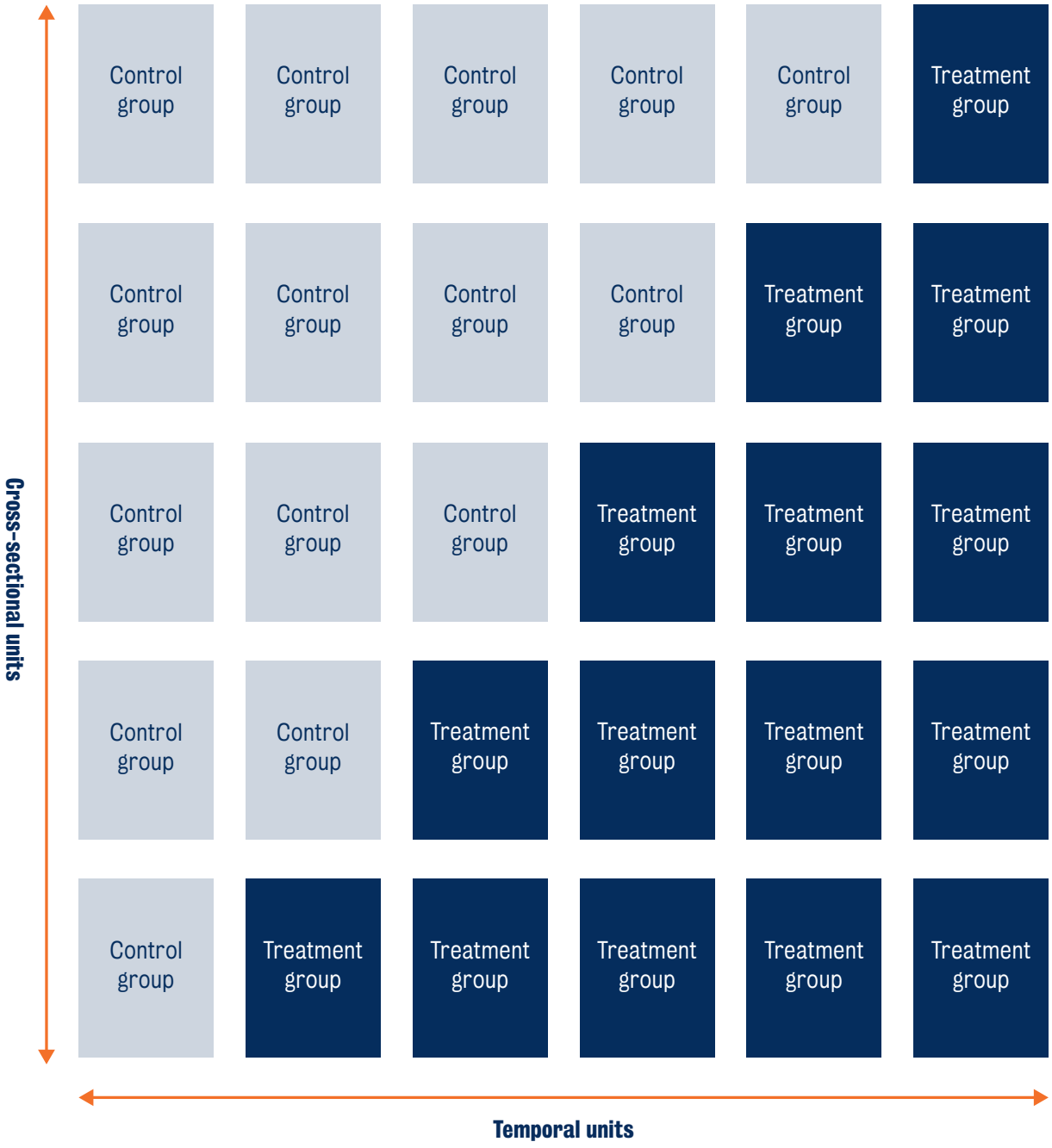
Introduction

Stepped-wedge randomised controlled trials (SWRCT) are an increasingly popular alternative to a parallel design, cluster randomised controlled trial, where the treatment and control groups are fixed across the duration of the trial. In short, a stepped-wedge trial starts with all clusters in the control state, then the intervention is introduced to randomly-selected clusters at intervals, until finally all clusters have been moved over to the treatment state. At every stage, outcome measurements are taken along with the treatment status of each cluster. By the end of the trial, all clusters will have passed from control to treatment, and thus each cluster is represented in its controlled and treated states. The reason behind the name 'stepped-wedge' is you can visualise the allocation of treatment by cluster as stair steps, resulting in a wedge-like shape representing the treatment duration for all the clusters across time, as seen in blue in Figure 1 overleaf.

Benefits of a stepped-wedge design

There are a few clear benefits to approaching a randomised trial in this way. First, you do not have to roll out an intervention all at once. If your intervention is very involved and/or expensive, and you are working under particular resource constraints (who isn't?), this can be hugely beneficial for the feasibility of your study. Second, your study can benefit in terms of statistical efficiency: since all participants and clusters play both treated and controlled roles, you are able to control both for period-effects, and for participant level effects, and may be more likely to detect an effect with a modest sample. And third, like a waitlist design, all clusters will eventually receive the intervention, which could make your trial more ethically acceptable in cases where the treatment presents a clear benefit (although there still may be well-founded objections to withholding treatment from later cohorts for an extended period).

Figure 1 A basic stepped-wedge design

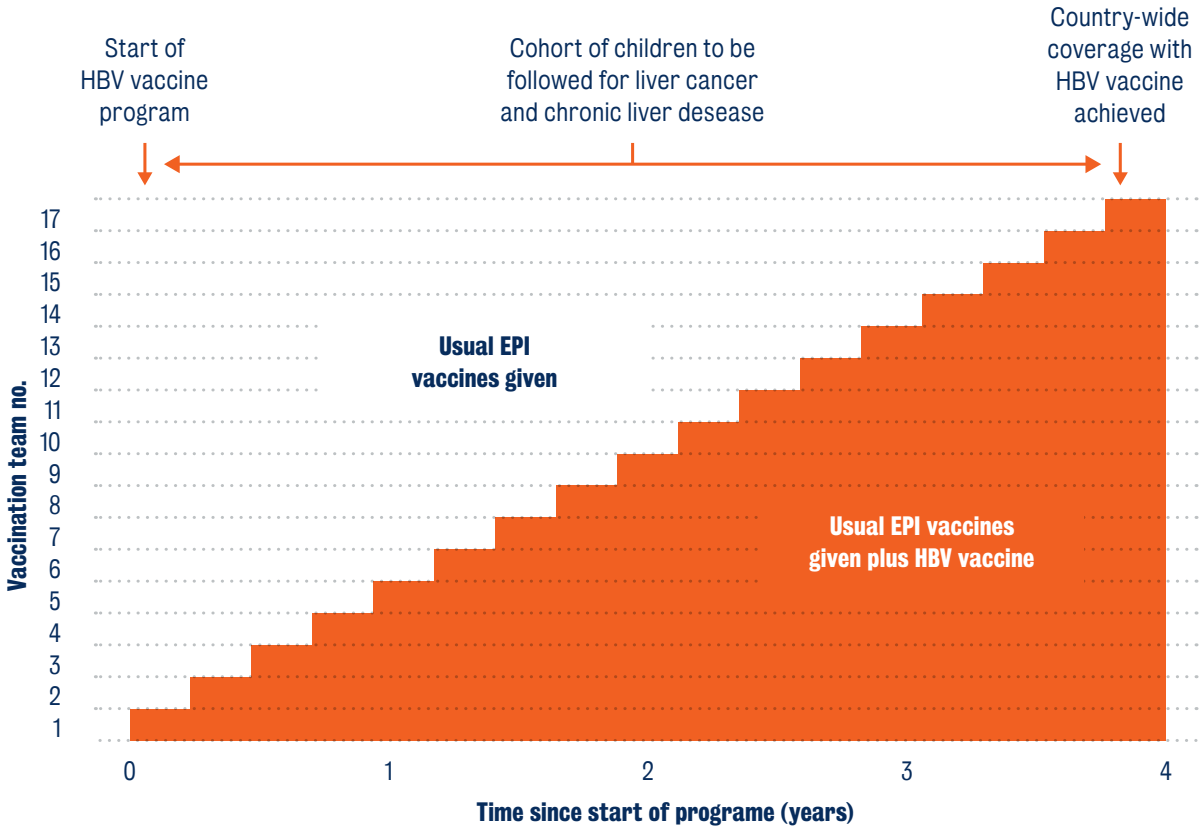


Example: the Gambia hepatitis intervention study

The first known example of using a stepped-wedge design is the Gambia hepatitis intervention study (Gambia Hepatitis Study Group, 1987; Hooper, 2020), which set out to understand whether adding a schedule of vaccinations against the Hepatitis B Virus (HBV) for newborn children in the Gambia would lead to decreased rates of liver cancer and chronic liver disease later in life. This study took place against the backdrop of a larger expanded programme of immunisation (EPI), a collaboration between the Gambian government and the WHO which aimed to expand childhood vaccination rates.

The clusters in this case were 17 immunisation teams that corresponded with geographical areas, and the participants recruited were newborn children born during the 4-year study starting in 1986. The treatment was the bundling of a HVP vaccine with other childhood vaccines that were being delivered routinely as part of EPI, while the controlled state was receiving childhood vaccines without the HPV component. The stepped-wedge for this trial is depicted in Figure 2 below, taken from the study design published in 1987.

Figure 2 The first stepped-wedge design. (Gambia Hepatitis Study Group, 1987)



The study team cited the following reasons for using a stepped-wedge:

- The HPV vaccine was expensive with limited availability
- The desire to have comparison groups within the same time period
- The logistical and ethical difficulties of randomising at an individual level
- The hope that by the end of the four-year period, the vaccination programme would be ready to provide HPV vaccination universally

Interestingly, this approach to a stepped-wedge design differs from the classical model: not all participants actually receive the intervention, and not all participants act as controls. Looking at Figure 2, you may notice that the first cluster (vaccination team 1) is never in the controlled state, and the final cluster (17) is never in the treatment state for the duration of the trial. It is also notable that while (nearly) all the clusters enter the treatment state eventually, the participants do not. It is only if the child is born after the relevant vaccination team has passed to the treatment state that they will receive the HPV vaccination schedule (eg a child born in the area covered by team 5 during the first year of the trial would always be in the control state). This choice not to vaccinate half of the participants in the trial was essential given the longer-term outcome measure of interest (liver disease and cancer), but it's worth noting that in this case, using a stepped-wedge design did not mean that all participants eventually received treatment.

When should a stepped-wedge design be used?

As other chapters have already made clear, there are many reasons why a traditional, individual randomised controlled trial might be objectionable or impractical, but equally there already are many options to overcome these objections and considerations: waitlist and crossover designs overcome the ethical challenges of withholding a treatment known to be beneficial, whereas cluster RCTs deal with logistical challenges of spillover and group-based treatments.

Given that we already have multiple treatment allocation designs, what value does stepped-wedge add, and when should it be used? Broadly, the stepped-wedge design can be considered when the following criteria are met:

- The trial does not rely on recruitment of individuals (ie treatment allocation is practical at the cluster level).
- The intervention is difficult to take away after it has been implemented, eg a training that has led to a change in practice, or the introduction of new. infrastructure or technology, or a new policy rolling out backed by legislation. The Gambia hepatitis example above introduces an intervention difficult to take away (ie a vaccine schedule) and was a new policy being rolled out anyway.

As discussed in the introduction, stepped-wedge designs are particularly appealing when:

- You are faced with scaling difficulties in implementing the intervention to all treatment clusters simultaneously. Since you can roll out the intervention incrementally, you can spread out the resources of your intervention delivery and research teams.

- The intervention is desirable; people want it. By using a design where what is being randomised is the order by which clusters receive treatment (rather than whether they receive treatment at all), you can overcome (some) objections over fairness, offer a transparent approach to a roll-out order for participants, and hopefully hedge against some attrition you would otherwise see amongst control units in a traditional parallel RCT.

- You're looking to maximise statistical efficiency, which is enabled by the fact that each cluster acts as both a treatment and control unit.

    - Cases where efficiency is not optimised by a SWT: this happens when the number of people recruited at each cluster in each step is relatively small, and when outcomes of people from the same cluster are not that much more similar than outcomes of people from different clusters.

> **Use of force trial**
>
> A stepped-wedge trial was used to evaluate a new training programme for police officers, carried out in partnership with the College of Policing and Avon and Somerset Constabulary. The new training programme aimed to reduce the extent to which police use force against suspects. A parallel randomised trial might have been possible, but it would not have been possible to train all of the officers in the constabulary – or even half of them – very quickly. Instead, they were assigned a different week to receive the new training, at random. This meant that while all officers in the constabulary were trained over the course of the 12 months (52 weeks), at any given point, only a random selection of them had been. In each week of the trial, we were able to compare the use of force by officers who had received the training, with the use of force by officers who had not yet received it. In this way, and because everyone was eventually treated, each time period of the data contained some treated and some control participants, and each police officer had some treated, and some control, time periods. We were also able to integrate historical data on use of force by police, allowing us to account for any differences that existed before the trial, as well as any seasonal variation in the use of force. The trial detected an 11 per cent reduction in the use of force by trained officers, compared to their untreated peers and their own past use of force. (Sanders et al., 2024)

Risks and challenges

Stepped-wedge trials can be complex to implement and come with their own set of risks and threats. First, as alluded to in the discussion of benefits, delayed access to a treatment with known benefits can be nearly as ethically problematic as withholding the treatment entirely. Second (and relatedly), lengthy waiting periods for receiving the treatment may lead to non-random attrition among the later clusters, which can both weaken the trial's statistical power and potentially introduce bias, if attrition is indeed non-random and these differences correlate with the outcome.

External, 'shock' type events (eg an economic downturn, a universal change to benefits, etc) are also more complex to account for in Stepped-wedge trials, and given that SWRCTs require certain spans of time to implement, there is more opportunity for shock events to occur. As there are fewer control clusters over time, we have fewer data points with which to construct a plausible counterfactual to the treatment condition, which makes it difficult to adjust for the effect of an external event on the outcome.

There are also numerous logistical considerations when undertaking a phased roll-out. You may need repeated training activities as each cluster transitions to treatment. Control arm clusters will need sustained engagement to prevent drop-out. These contribute to an increased workload for the intervention and research teams over time, as additional clusters join the intervention arm.

Finally, it is worth reflecting that stepped-wedge designs are still relatively new. Given the complexity of a stepped-wedge, more time will need to be dedicated to laying out the details of the roll out periods, the size and division of clusters, how external events will be captured and adjusted for, etc, and there will be relatively limited literature to guide these decisions compared with other designs. That said, there are now published CONSORT guidelines for reporting on stepped-wedge trials, which go into some detail on the methodological complexities and offer baseline reporting recommendations (Hemming et al., 2018).

Study design and analysing data from a stepped-wedge design: accounting for time and clusters

One could, in theory, analyse data collected from stepped-wedge trials using the same approach as cluster randomised controlled trials: estimate effects for individuals, controlling for known covariates and using cluster-robust standard error calculations. This approach however would ignore the role of time, which has the potential to confound the intervention-outcome relationship. As time passes, trends and events external to your study can have differential, interactive impacts on each cluster, depending on when that cluster crosses over from control to treatment. Time thus must be accounted for, both at the design stage and analysis.

Another factor for consideration in a stepped-wedge is the size and composition of the clusters. As mentioned above, stepped-wedge trials can yield better statistical efficiency (ie power) than traditional cluster RCTs, but don't treat this as a given. Generally speaking, stepped-wedge designs outperform CRCTs in terms of power in two cases: first, where the intra-cluster correlations are relatively high (see Chapter 2 on cluster randomised trials for an explanation on ICCs) within clusters ie participants systematically differ from those in other clusters; and second, where clusters are relatively large and there are fewer clusters (Hemming et al., 2015). As we discussed in the relevant chapter, a high ICC diminishes statistical power, and a high ICC paired with few clusters means fewer opportunities to detect differences between treatment and control that are attributable to the intervention. Stepped-wedge designs mean we can collect data from all units in both their treatment and controlled states, increasing the available data for analysis.

Power analysis

Conducting a power analysis for a stepped-wedge design, as may be inferred already, must account for the ICC in addition to the typical inputs of anticipated effect size, variation of inputs/outputs, and required statistical power. So far, so good – we already know how to incorporate ICC into power calculations through the calculation of the design effect (see the chapter on CRTs). The added complexity we need to contend with is the impact of time: in particular, heterogeneous treatment effects as a result of time (an important factor to account for), and the likelihood of intra-cluster correlation rates decaying over time (less big, since lower ICCs are generally viewed as beneficial). The literature on these points gets really complex, really quickly (see Li et al., 2020 for more information), so in sum for the first point: heterogeneous treatment effects can be understood as an interaction between treatment and time. Put another way, a treatment can be more or less effective, depending on environmental or participant-level factors at play relating to time. To the second point on ICCs and time, consider how members of a cluster may share things in common at the beginning of a study, but then as time wears on and entropic forces intervene, members start to drift and resemble each other less than they did at the beginning.

For stepped-wedge designs, one can calculate the required sample size by multiplying the standard individual-level randomised design sample size by the design effect as would be done in parallel randomised trial designs. A simple step-by-step guide is provided below following from Hemming & Taljaard, (2016), on how to conduct power calculations for SW designs. In this example, clusters are assumed to be equally sized.

Step 1: Calculate the required sample size (N1) given the desired effect size using the formula for a parallel individual randomised trial.

Step 2: Decide on a desired intra-cluster correlation (ICC), the size per cluster (M) and the number of steps (t) of the SW design. Then calculate the SW sample size per cluster per step (m) using the formula:

$$m = \frac{M}{(1+t)}$$

Step 3: Calculate the SW design effect by plugging in the ICC, m and t values:

$$DE_{sw} = (t+1)\frac{1+ICC(tm+m-1)}{1+ICC\left(\frac{tm}{2}+m-1\right)}\frac{3(1-ICC)}{2\left(t-\frac{1}{t}\right)}$$

Where ICC is the intra-cluster correlation, m is the average cluster size per time point and t the number of time points.

Step 4: Multiply the individual-level randomised sample size (N1) by the design effect to get the adjusted SW total sample size (N):

$$N = DE_{sw} \cdot N1$$

The number of clusters (k) can then be inferred using the following relation:

$$k = \frac{N}{m(1+t)}$$

For example, assuming an effect size of ES=0.2 with a std. deviation of 1, significance level of 0.05 and power 80%, an individual-randomised trial would require N1=788 participants. To adjust for a SW design, assuming an ICC=0.01, a cluster size of M=30 and 4 steps we would have:

$$DE_{sw} = (4+1)\frac{1+0.01(4\cdot 7.5+7.5-1)}{1+0.01\left(\frac{4\cdot 7.5}{2}+7.5-1\right)}\frac{3(1-0.01)}{2\left(4-\frac{1}{4}\right)} = 2.22$$

$$N = 788 \cdot 2.22 \approx 1749$$

Or approx. 11 clusters per step with around 8 units per cluster per step).

For the sake of expediency, there are some online power calculators to come to our rescue in calculating sample size. This example, built in Shiny R, allows the incorporation of varying cluster sizes and numbers of steps, as well as ICC and ICC rate decay over time: https://clusterrcts.shinyapps.io/rshinyapp/ (Hemming & Kazsa).

Estimating effects

In stepped-wedge trials treatment effects can be estimated using mixed effects modelling (MLM) as proposed by Hussey and Hughes (2007). MLMs are well suited for SW designs since they can account both for the variation within clusters (at the level of the individual unit) and variation at the cluster-level via the inclusion of random intercepts. They can also be extended to account for heterogeneity stemming from time effects via the inclusion of time random slopes (Hooper et al., 2016). Further to this the covariance structure between random slopes and intercepts a can be modified (exchangeable, autocorrelational, etc.) depending on the design characteristics of the SW trial (cross-sectional, cohort).

Effect sizes like Cohen's G can then be estimated by standardising the raw treatment effect by the total variance of the model.

Conclusion

The stepped-wedge trial design offers a flexible and pragmatic approach to evaluating interventions, particularly in scenarios where a traditional randomised controlled trial is not feasible. Its ability to balance logistical constraints, ethical considerations, and statistical efficiency makes it a valuable tool in the researcher's arsenal. However, as highlighted in this chapter, the design is not without its limitations and complexities.

Despite its advantages, the stepped-wedge design requires meticulous planning given the large number of design permutations and a carefully matched analytical approach. Researchers must also be aware of the possible bias that can stem from failing to account for sources of confounding, all of which add layers of complexity to the study's design and analysis. Similarly, and equally importantly, power calculations must also incorporate these factors to ensure the trial's robustness and interpretability.

In conclusion, while the stepped-wedge design holds significant promise, particularly in the context of scalable, desirable, and irreversible interventions, its implementation demands careful consideration of both methodological and logistical challenges. As the field continues to evolve and the literature grows, this design will likely become an increasingly important option for researchers faced with complex intervention scenarios.

Part 2
Tools of the trade

Part one has provided an overview of the main kinds of trial and how they are conducted, as well as their strengths and weaknesses.

In Part two, the focus changes to some more general questions about how to run trials more concretely. Over the next several chapters, we will cover some of the key things you need in order to actually put a trial into practice, As a result, these chapters are more technical in nature, but throughout we have attempted to include the intuition behind some of the topics that are discussed.

Included within part two are chapters on:

- Trial protocols – the essential document produced for each trial which fixes the trial design and ensures the robustness of statistical analysis.
- Sample size calculations and statistical power – the key calculations that help you calculate how large your trial needs to be
- Missing data – how to handle a situation where either participants haven't made it to the end of the trial, or some of their data are missing — for example if they didn't provide some baseline data – which can be one of the major risks to the validity of a trial.
- The interaction between statistical power and inequality, and how we can run trials that square the circle, letting us better understand the equity impacts of our trials.
- Issues of multiple comparisons – what happens when we design studies with multiple outcomes measures; multiple subgroups of interest; or more than two conditions, and how we can ensure the robustness of designs like this.
- Consent vs assent in randomised controlled trial designs, and the practical vs ethical trade-offs of opt in vs opt out consent.

Chapter 5. Protocolisation

Once you have decided to run a trial, the most important thing that needs to be produced is a trial protocol. A trial protocol is a document that describes what the trial is aiming to learn, and how it aims to go about it. Producing a protocol can be an enormous amount of work, but doing it before the trial begins will improve the rigour of the research, and save time later.

What is a protocol for?

There are three main purposes to a trial protocol. These are;

1. To describe ahead of time all of the details of the trial, including what the intervention is; the intended sample size; how randomisation will be conducted; what outcomes will be collected, and how they will be analysed. This in turn allows for scrutiny and peer review ahead of the trial being run, in a way that allows any issues to be worked out.

2. To allow someone coming into the project part way through to understand the nature of the trial — for example, if a new project manager, or a policy official, arrives midway through, they can read the protocol and have a good sense of what the trial is, and what it is and isn't for.

3. To constrain the decisions of the evaluators and any other parties to the evaluation when the data are collected, and to ensure that the analysis and reporting take place as intended.

These three benefits are inseparable and inter-related, but nonetheless distinct. A project without a protocol is more likely to contain flaws; is more likely to experience struggles related to changes in personnel; and is going to be less robust. In order to ensure that a trial meets its goals, the protocol should be published in advance of the trial launching, or at least in advance of the data being analysed.

Why protocolise?

It might seem that protocolising a trial in this way is a natural thing to do. However, in many elements of social science and social policy research, formal protocolisation, and the attendant pre-registration, was relatively uncommon. Where protocols were produced, they were often internal documents — and hence subject to change, and difficult to use for accountability purposes.

We need to protocolise so that commissioners; evaluators; and the people developing and delivering an intervention have a shared understanding of what the intervention is, and what the outcomes are that are being tested. This is important for avoiding confusion and disagreement later. Specifically, if an evaluation concludes that something doesn't have its intended effects, an agreed research protocol limits the room at the end of the programme for arguing that the evaluation was in fact measuring the wrong thing.

We also need to protocolise to ensure that research is conducted in a robust way. A protocol details the outcome measures to be collected, and which are the primary, secondary, or exploratory outcomes — essentially ranking them in terms of importance. Doing so allows us to say, after the fact, how important the findings are — preventing anyone from claiming that a secondary measure for which there was a significant result was the intended main outcome all along.

This tying of hands also applies to the statistical analysis to be conducted. Even in a simple trial, researchers and evaluators have a range of options available to them in terms of how to conduct the statistical analysis. This might include using different sorts of statistical tests; removing some observations for being outliers; changing the distribution of those variables (by taking logs, for example), or changing the way in which variables are combined in order to make an composite variable; what variables are controlled for in their analysis. They also have choices around whether to look at sub-group analyses (whether there was an effect for a particular group, for example women), and how to approach these analyses. Finally, there are choices about who we class as 'treated' — whether we make use of intention to treat analysis (in which someone's treatment status is based on the group they were assigned to), or a compliance analysis (in which their status is based on whether they complied with the treatment). These decisions are called "researcher degrees of freedom".

Many of these decisions are justifiable — we might legitimately be interested in effects of a treatment on women, and there are certainly circumstances in which outliers should be removed. The problem emerges when we choose which decisions to make once we have the data, based on the results they produce. If we get the data and analyse the results, trying different strategies, we can then pick the ones that give us the results we want. Even with an independent evaluator, this still provides a natural bias towards finding effects that are positive and significant. This process — of trying different approaches until we get the result we want — is called P-hacking (see box for an example), and is the equivalent of cheating at darts by throwing a dart at the wall and painting the dartboard around where your darts landed to give the impression that you hit the bullseye. There is clearly a difference between setting out to discover whether an intervention has differential effects on women, and claiming after you find a significant effect for women that this was your objective all along.

Elements of a protocol

There are a range of different elements of a protocol that you might want to include, and the exact mix you choose will depend on the field you're working in. However, there are some key elements that need to be included — which are listed below.

- The name of the intervention
- Details of the evaluators and delivery partners
- The evaluation research questions
- Details of what the intervention is and how it will be implemented.
- The study design — how randomisation will be conducted, whether the trial is individually randomised, cluster randomised, and so on.
- A study timeline — what will happen and when
- Participant details including who is eligible and who is not
- The outcome measures, how they are collected, and whether they are primary, secondary, or exploratory.
- How statistical analysis to be conducted, including detailed analytical specifications.
- Any compliance analyses
- A trial protocol template can be found in Annex 1 of this book.

You may also wish to include details of an implementation and process evaluation, or an economic evaluation, which will identify how the intervention is implemented, whether it is implemented with fidelity, how people feel about the intervention, and whether the intervention represents value for money. Although these are integrated parts of an RCT, they are not *essentially* part of an RCT.

You may also wish to include how you will handle compliance analyses; how you will handle missingness in the data, and attrition from the sample. However, these might also be specified in analytical guidance, for example that produced by many what works centres, in which case it might not need duplicating in your protocol.

Details about the intervention

For producing a description of the intervention you are evaluating, it can be useful to use a standardised framework that allows you to understand the important elements of an intervention at a glance, and to compare different interventions tested in different trials. This can be achieved using a theory of change, which outlines what the intervention is and how it is posited to work, or by using a framework like the Template for intervention description and Replication (TIDier). Below, we show an example of both for a trial providing personalised budgets to people with people living in temporary accommodation, which is being run by the Centre for Homelessness Impact (Sanders et al, 2024).

Figure 1 Intervention at a glance

Situation	Financial hardship is a contributing factor to both initial spells of homelessness and the likelihood of experiencing spells. Many existing schemes do not offer enough money to make meaningful change. People with experience of rough sleeping may benefit from having support from a caseworker while maintaining agency in determining what best to spend money on. Many interventions or policies deny agency to people with experience of homelessness.			
Aims	The Greater Change Personalised Budgets intervention aims to provide targeted financial support, in a volume that can make a meaningful difference, to people with experience of homelessness, with a goal of reducing homelessness			
Process			**Impact**	
Context	**Activities**	**Outputs**	**Outcomes**	**Impact**
Participants are people who have experienced rough sleeping, and who are now in temporary accomodation, who have less that £4000 in savings. These participants have a pre-existing relationship with a charity partner, and have a case worker. Charity partners are organisations that work with people in temporary accomodation which do not currently routinely offer budget support of £400 or more. Greater Change's infrastructure: referral processes; and the funding of ~£4000 per person	Participants will be referred by their case worker. Participants, case worker and Greater Change will complete the referral process. The amount of money, and what it is to be spent on, is agreed between Greater Change, charity partner, and the participant. Greater Change transfers the money within four weeks to the charity partner. The case worker spends the money as agreed.	The purchase of whatever the money is agreed to be spent on, which can be a wide range of things (articulated in the research protocol). Greater understanding of move-on plan between participants and case workers.	Participants are in a more **financially secure position** due to the money being spent. Participants are in a more **secure housing** position that they would otherwise have been. Participants have **higher subjective wellbeing** than they otherwise would have. Participants are more, or no less, likely to be **in employment**. People are no more likely to spend moeny on illicit drugs (adverse outcome).	People are less likely to experience homelessness, and any experiences of homelessness are **rare, brief and non-recurring**.

TIDieR Framework

Figure 2 TIDieR Framework

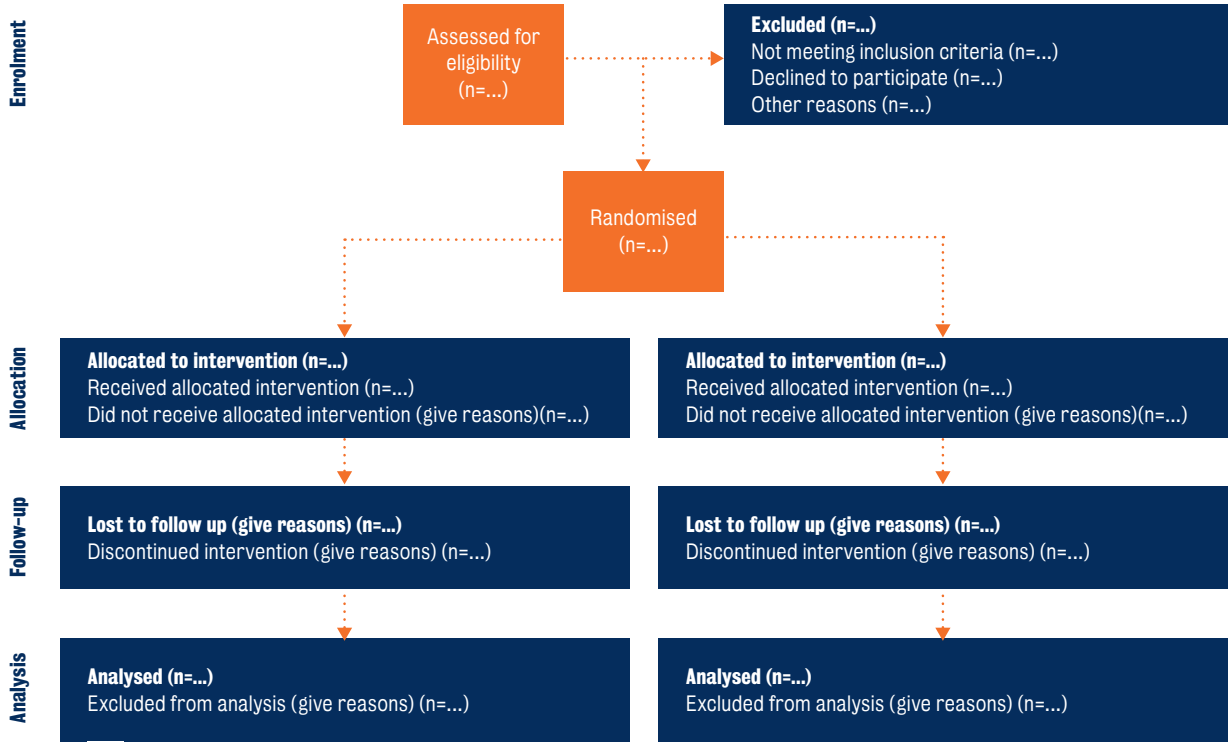
Name	Greater Change Personalised Budgets Intervention
Why	Poverty and low income are major contributors to people's risk of homelessness, and their inability to escape homelessness and its consequences. There is therefore a an argument for more financial support being given to people experiencing homelessness. However, people experiencing homelessness can benefit from support in deciding how that money should be spent, and a large cash transfer might put them at risk of exploitation. The Greater Change Personalised Budgets Intervention therefore works with people experiencing homelessness, and with a charity worker who supports them, to identify a concrete need that a personalised budget can be spent on; helping ensure that the money is spent appropriately in a way that maximises benefit to the recipient, and that the risk of harm is significantly reduced.
Who (recipients)	People with previous experiences of homelessness, currently housed in temporary accommodation. Participants will be 18+ and have an existing relationship with a case worker from an eligible charity partner. Charity partners are eligible if they work with people in temporary accommodation, and if they do not currently provide substantial financial support (the modal £100 interval provided by any existing financial support scheme is less than £400).
What (materials)	A payment of a specified amount agreed between Greater Change, the participant and their caseworker, provided to the case worker to make the purchases agreed. The amount provided, and what it is spent on, will vary — a list of types of things that the intervention is spent on can be found in the appendices. Average values of goods/services provided with the personalised budget is expected to be around £4,000.
What (procedures)	Participants are referred to Greater Change via a referral form, A discussion takes place between the case worker; Greater Change; and the participant, to decide the best use of the personalised budget. When this is agreed, Greater Change will transfer the budget to the partner charity within 4 weeks, who will spend the money as agreed.
Who (provider)	The intervention will be delivered by Greater Change, in partnership with charity partners.
How (format)	Money agreed will be transferred to the charity partner's bank account and used to purchase what was agreed.
Where (location)	Across England (exact locations tbc)
When and how much (dosage)	Payments will be made within 4 weeks of being agreed. It is anticipated that average payments will be around £4,000, but they may exceed this where necessary.
Tailoring	The personalised budget planning conversation will be tailored to the participant's individual aspirations, goals and requirements.
Control condition	Control participants, as with treatment-arm participants, continue to be able to access all other support that they are entitled to (e.g. universal credit, rent assistance). Control participants will not be allocated a personalised budget but are entitled to all other support.

Describing your trial

As well as describing the intervention, you also need to provide a description of the trial design itself. Whilst this can and should be done using prose, it can also be helpful to make use of a standard way of visualising and reporting trials. The most widely used such template is a CONSORT diagram, which is derived from the CONSORT Guidance. These diagrams are initially designed for medical contexts, but can be readily adapted to a policy context, and there are various versions of them (see the bibliography of this chapter), for different types of trial). However, the core CONSORT guidance requires the reporting of the following elements, with the recommended flow diagram beneath (Moher et al, 2010):

- Title and abstract – Identification as a randomised trial in the title (1a)
- Introduction – Scientific background and explanation of rationale (2a); Specific objectives or hypotheses (2b)
- Methods – Trial design (3a, 3b); Participants (4a, 4b); Interventions (5); Outcomes (6a, 6b); Sample size (7a, 7b); Randomisation (8a-12b)
- Results – Participant flow (13a, 13b); Recruitment (14a, 14b); Baseline data (15); Numbers analysed (16); Outcomes and estimation (17a, 17b); Ancillary analyses (18); Harms (19)
- Discussion – Limitations (20); Generalisability (21); Interpretation (22)
- Other information – Registration (23); Protocol (24); Funding (25)

Figure 2 Consort



Exploratory analyses

One frequent criticism of protocolisation and pre-registration is that it stifles exploration, by limiting analyses to those included within the protocol. This needn't be the case — there is no reason why researchers working on a randomised trial might not conduct exploratory analyses that have not been pre-registered. Where these emerge throughout the trial — for example when implementation of the intervention leads to the discovery of a particular group that seems to engage with the intervention more than anticipated, then a protocol addendum might be appropriate before analysis is conducted.

When surprising subgroup effects, or effects on unexpected outcomes emerges at the stage of analysis — this is absolutely fine, this kind of discovery is an important part of research. What matters is that the nature of this discovery is presented transparently. When a finding emerges from exploratory analysis, it should be flagged as such, and we shouldn't pretend that it was one of the hypotheses held at the beginning of the research project — which is what a protocol seeks to prevent.

What is P-hacking?

One thing that protocols seek to prevent is p-hacking. This term, popularised by Simmons et al (2011), refers to the ability of researchers to run a wide range of analysis from a single experiment, in a way that increases the likelihood that they find the result they want — usually a significant and positive effect of an intervention. This should be distinguished from making decisions at the start of a trial that maximise the statistical power of the trial.

P hacking can be done in a number of ways, but the common element is by running different analyses, and choosing to report — or highlight — only those that confirm your hypothesis, or which are otherwise interesting. In Simmons et al's paper, they create a novel example which demonstrates p-hacking. They recruit student participants, and have them either listen to no song, to listen to When I'm 64 by the Beatles, or another song. They collected a range of characteristics about participants, and by excluding the third condition (with the second song), and by running lots of analyses, they were able to find a test which showed statistically that listening to 'When I'm 64' makes you younger. Clearly, this cannot be true — because your age cannot be affected by being made to listen to a particular song — but by testing repeatedly until they found the desired effect, Simmons et al were able to 'prove' that it does. Although this example is trivial, the same can occur with policy trials where the analytical approach isn't pre-specified. This, at least in part, explains the large disparity in effect sizes detected between independently evaluated studies (which have smaller effects and more null results on average), and those where the research is conducted by the intervention developers (which have large effect sizes), because of the disparity in the motivation to p-hack.

References

CONSORT guidance: Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gøtzsche, P. C., Devereaux, P. J., ... & Altman, D. G. (2010). CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ*, 340.

CONSORT Non-inferiority: Piaggio G, Elbourne DR, Pocock SJ, Evans SJW, Altman DG, for the CONSORT Group. Reporting of noninferiority and equivalence randomized trials. Extension of the CONSORT 2010 statement. JAMA. 2012; 308(24): 2594-2604.
PMID: 23268518

CONSORT Cluster: Campbell MK, Piaggio G, Elbourne DR, Altman DG; CONSORT Group. Consort 2010 statement: extension to cluster randomised trials. BMJ. 2012;345:e5661.
PMID: 22951546

CONSORT Pragmatic Trials: Zwarenstein M, Treweek S, Gagnier JJ, Altman DG, Tunis S, Haynes B, Oxman AD, Moher D; CONSORT group; Pragmatic Trials in Healthcare (Practihc) group. Improving the reporting of pragmatic trials: an extension of the CONSORT statement. BMJ. 2008;337:a2390.
PMID: 19001484

TIDieR: Hoffmann T, Glasziou P, Boutron I, Milne R, Perera R, Moher D, Altman D, Barbour V, Macdonald H, Johnston M, Lamb S, Dixon-Woods M, McCulloch P, Wyatt J, Chan A, Michie S. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. BMJ. 2014;348:g1687.
PMID: 24609605

CONSORT for pilot and feasibility trials: Eldridge SM, Chan CL, Campbell MJ, Bond CM, Hopewell S, Thabane L, Lancaster GA; on behalf of the PAFS consensus group. CONSORT 2010 statement: extension to randomised pilot and feasibility trials. Pilot and Feasibility Stud. 2016;2:64.
PMID: 27965879

CONSORT for within person randomised trials: Pandis N, Chung B, Scherer RW, Elbourne D, Altman DG. CONSORT 2010 statement: extension checklist for reporting within person randomised trials. BMJ. 2017;357:j2835.
PMID: 28667088

CONSORT Equity: Welch VA, Norheim OF, Jull J, Cookson R, Sommerfelt H, Tugwell P; CONSORT-Equity and Boston Equity Symposium. CONSORT-Equity 2017 extension and elaboration for better reporting of health equity in randomised trials. BMJ. 2017;359:j5085.
PMID: 29170161

Stepped wedge cluster randomised trials: Hemming K, Taljaard M, McKenzie JE, Hooper R, Copas A, Thompson JA, Dixon-Woods M, Aldcroft A, Doussau A, Grayling M, Kristunas C, Goldstein CE, Campbell MK, Girling A, Eldridge S, Campbell MJ, Lilford RJ, Weijer C, Forbes AB, Grimshaw JM. Reporting of stepped wedge cluster randomised trials: extension of the CONSORT 2010 statement with explanation and elaboration. BMJ. 2018;363:k1614.
PMID: 30413417

CONSORT for multi-arm parallel-group randomised trials: Juszczak E, Altman DG, Hopewell S, Schulz K. Reporting of Multi-Arm Parallel-Group Randomized Trials: Extension of the CONSORT 2010 Statement. JAMA. 2019;321(16):1610-1620.
PMID: 31012939

CONSORT 2010 extension to randomised crossover trials: Dwan K, Li T, Altman DG, Elbourne D. CONSORT 2010 statement: extension to randomised crossover trials. BMJ. 2019;366:l4378.
PMID: 31366597

CONSORT Factorial: Kahan BC, Hall SS, Beller EM, Birchenall M, Chan AW, Elbourne D, Little P, Fletcher J, Golub RM, Goulao B, Hopewell S, Islam N, Zwarenstein M, Juszczak E, Montgomery AA. Reporting of Factorial Randomized Trials: Extension of the CONSORT 2010 Statement. JAMA. 2023;330(21):2106-2114. PMID: 38051324

CONSORT-Surrogate: Manyara AM, Davies P, Stewart D, Weir CJ, Young AE, Blazeby J, Butcher NJ, Bujkiewicz S, Chan AW, Dawoud D, Offringa M, Ouwens M, Hróbjartssson A, Amstutz A, Bertolaccini L, Bruno VD, Devane D, Faria CDCM, Gilbert PB, Harris R, Lassere M, Marinelli L, Markham S, Powers JH 3rd, Rezaei Y, Richert L, Schwendicke F, Tereshchenko LG, Thoma A, Turan A, Worrall A, Christensen R, Collins GS, Ross JS, Taylor RS, Ciani O. Reporting of surrogate endpoints in randomised controlled trial reports (CONSORT-Surrogate): extension checklist with explanation and elaboration. BMJ. 2024;386:e078524. PMID: 38981645

Sanders M, Hirneis V & Vallis D. (2024). Trial Protocol: Personalised Budgets Randomised Controlled Trial. Centre for Homelessness Impact. https://cdn.prod.website-files.com/646dd81ef095aa13072c44e0/66cee91d656ca3ee3d51f78c_T%26L%20Protocol%20-%20Personalised%20Budgets%20Trial%20-%20MHCLG%20APPROVED.pdf

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2018). False-positive citations. Perspectives on Psychological Science, 13(2), 255-259.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological science, 22(11), 1359-1366.

Chapter 6. Sample size and power

What is statistical power?

In plain terms, statistical power helps you figure out how large your sample must be in order to detect the effect of your intervention (if there is an effect). Statistical power relates to the concept of a type II error, which is the failure to reject the null hypothesis when there is in fact an effect (also known as a false negative; in contrast, type I errors are false positives, or detecting an effect where there is none). Statistical power is expressed in terms of a probability, thus it will be a value between 0 and 1, which indicates the probability of avoiding a type II error.

Practically speaking, statistical power is highly important to consider *ex ante*, at the planning stage of an evaluation trial. Specifically, statistical power is calculated via a **power analysis**, which incorporates key elements of the trial in order to project how large of a sample one would need to find an effect of an intervention. This is important to get right, as a study with insufficient power may indicate that an intervention is ineffective (when in fact it is not), which has negative implications both in terms of wasted trial resources and in terms of contributing to evidence-based policy. Underpowered studies are also more likely to experience type S errors – when the sign of the effect is estimated incorrectly – and type M errors – where the magnitude of the effect is estimated incorrectly. Taken together, these dramatically reduce the confidence we can have in underpowered studies' findings.

Statistical power can be a high stakes game, and it is not helped by the fact that power analysis itself can be an imprecise and complex process, especially when there is a complex intervention or randomisation protocol. This is in part due to the ambiguity surrounding different values that power analysis incorporates, particularly the anticipated effect size of the intervention. The reader now may well be thinking: How are we meant to predict the effect size *before* implementing the trial and collecting the data? Isn't that the whole point of the trial? Well, exactly. Power analysis requires a degree of guesswork and informed intuition, and is highly sensitive to context, discipline, and methodological approach. That said, we do not have to fly blind: there are rules of thumb and strategies that can take some of the guesswork out of it. In this chapter, we will make the case for why statistical power and power analysis are important elements of any quantitative analysis, then cover the practicalities of running a power analysis, and finally turn to threats to statistical power and potential complications.

Why do we calculate power?

Answering this question is at once obvious and also somewhat opaque. Of course it is helpful to know how many people/schools/local authorities etc we should sample in order to test an intervention. But why not just gather as large a sample as possible, and if we are lucky, we will find an effect? It comes down to two main reasons: cost (efforts both to minimise cost by limiting the sample *and* to avoid wasted resources by launching an unpowered trial), and ethics (respecting equipoise).

Cost

As with most things in life, choices in research can be understood as a series of trade-offs: optimising in one area usually means compromising in another. For most studies that involve participant recruitment (ie pretty much all RCTs), there is a trade-off between the number of participants – or units of analysis – that can be recruited for a single study, and the volume of research that a researcher or team can possibly do (Sanders et al., 2020). Recruitment, treatment allocation, trial management, and data collection almost always incur costs (money, time), and that cost varies depending on the size of the trial. With every additional unit of sample size, the researcher increases the chances an effect of a given size will be detected, but this probability increase is not linear: at some point, the benefit of adding more and more units will eventually drop close to zero. Power analysis helps us estimate where that sweet spot is, where we have enough units to reliably detect a given effect size, but not so many that we potentially waste study resources for ever-diminishing returns.

Ethics

Another unignorable reason to do power analysis is ethics. While a researcher may be completely convinced that an intervention works before doing the trial (it happens to the best of us), the whole reason we do trials is because there is genuine uncertainty around an intervention's effects, both in terms of magnitude and direction – if an effect exists at all. In clinical research, this state of uncertainty about effect is often referred to as equipoise (discussed in more depth in Chapter 1), which is a baseline requirement for all trials of therapeutic methods. This is because an intervention meant to improve outcomes can sometimes do the opposite: we run the trial to be sure, but this uncertainty behoves us to limit the number of participants to a number needed to find an effect.

Even if an intervention is very low-risk, there are still often costs to participants, in the form of time, commitment of resources, travel, etc. Researchers have an ethical obligation to limit burdens placed on participants, particularly if those burdens can outweigh the potential benefit of the research (Sanders et al., 2020).

Now that we are fully motivated to conduct useful power analyses, let's turn to the basics of how to do them.

Power analysis: the basics, then introducing complexity

The basics

Power analysis refers to the process of estimating different values relating to the statistical power of a given study, and takes into account five things:

1. The real size of the effect.

2. The variation of X (ie the range of treatment; a two-arm trial would have a simple binary variation of 1 or 0).

3. The variation of Y (ie the range of measured data for our outcome of interest).

4. Statistical power (ie the probability of correctly rejecting the null hypothesis if an effect exists; 80 per cent is often conventionally used).

5. The sample size.

You will need any four items from the list in order to estimate the remaining fifth item. In most cases, a researcher will be interested in finding #5, the right sample size to aim for in testing an intervention or running an analysis.

Let's consider a simple example to illustrate the five elements of power analysis. Let's say a researcher is testing the effect of a weekly SMS message to parents with children who are behind on their reading skills that reminds them of the benefits of reading to their children. Based on review of similar studies and her own intuition, the researcher anticipates that the SMS message campaign will lead, on average, to a five-point increase (on a 100-point scale) on a standardised reading assessment for children in the treatment group. Let's say she's comfortable with an 80 per cent probability of rejecting the null hypothesis correctly (#4; 80 per cent probability is commonly used). She knows the variation of X is binary (#2: SMS treatment = 1, 0). She also can know the variation of Y, based on the baseline reading scores of both the treatment and control groups, pre-treatment (#3). So, now that she has #1-4, she can now project the sample size she should aim for.

But how? There are a few ways to approach this, with the simplest being just taking advantage of the many free, online calculators such as this one (HyLown Consulting, 2022).

We can also do power analysis using analytical software such as R. Packages like pwr (Champely, 2020) offer the basic functionality needed for running power analyses with varying parameters. For most purposes, using these packages or an online calculator is completely fine. But if you want to understand the logical steps at work in these package functions, read on below.

To conduct a power analysis 'by hand' (but with the assistance of software), we will go through the following steps:

1. Generate a simulated dataset with the relevant properties (including #1, #2, and #3 above).

2. Regress Y on X using the simulated dataset, extract the p-value for the X coefficient, and determine whether the p-value is significant (ie less than 0.05, conventionally).

3. Loop steps 1-2 (for example 200 times) and store the reported significance of each model (eg signif = [FALSE, TRUE]).

4. Return the mean of all the reported significance: this is our statistical power, or the probability of correcting rejecting the null hypothesis.

In order to find the minimum viable sample size *given* the statistical power, we add the following steps:

1. Create a list of sample sizes to try.

2. Loop the above steps, taking each listed sample size in turn.

3. Store results in a table with columns sample size and statistical power, and see at which sample size we cross 80 per cent power.

We can benefit from creating a function that completes steps 1-4 and takes effect and sample size as parameters. We will use packages from the R tidyverse (Wickham et al., 2019) and the broom package (Robinson et al., 2023).

```
#Ensure required R package libraries are installed. Setting the seed ensures
replicable results.

library(tidyverse)
library(broom)

set.seed(42)

#This function will generate 500 simulated datasets using a for-loop with a
mean of 60 and a sd of 15, run a regression model for each, and store the
significance in sig_results

my_power_function <- function(effect, sample_size) {
 sig_results <- c()

 for (i in 1:500) {
  tib <- tibble(
  X = sample(c(1,0), sample_size, replace = TRUE, prob = c(.5,.5))
) per cent> per cent
 mutate(Y = effect*X + rnorm(sample_size, mean = 60, sd = 15))

  # Run the analysis
  model <- lm(Y ~ X, data = tib)

  # Get the results
  sig_results[i] <- tidy(model)$p.value[2] <= .05
 }

 sig_results per cent> per cent
  mean() per cent> per cent
  return()
}
```

Then, we can use this function to explore the minimum required sample size in steps 5-7:

```
power_levels <- c()

sample_sizes_to_try <- c(75, 100, 125, 150)

for (i in 1:4) {
 power_levels[i] <- my_power_function(5, sample_sizes_to_try[i])
}

# Where do we cross 80 per cent?
power_results <- tibble(sample_size = sample_sizes_to_try,
            power = power_levels)
power_results
```

Understanding the underlying logic of power analysis can be helpful as we introduce complexity to the picture, including clustered randomisation and complex interventions, which we will turn to shortly. However, for most users, using off-the-shelf software is likely to be sufficient.

A note on effect size: existing evidence, rules of thumb, repetition, and caution

As noted before, estimating the probable effect size is an essential part of determining the minimum required sample size. How do we go about that? While there are rules of thumb we can call upon, estimating the probable effect size can be just as much art as science, and as with many things in life, there are trade-offs to consider. A sensible place to start is by looking at the **existing evidence**: are there other trials testing a related intervention with a similar target population? What effect sizes did they find? In some cases, others have already done much of the work for you by undertaking **meta-analyses** across trials done in a particular sector. Works by Hattie (2008), Lipsey & Wilson (1993) and Wiliam (2008) review the educational intervention space and offer their own descriptive analyses of average effect sizes across different intervention types and trial designs. These average effect sizes are often expressed in terms of standard deviations to create comparability across many different measurements and outcome distributions.

Cohen (1998) proposed a set of guidelines for interpreting effect sizes using **Cohen's d**, a statistical measure which quantifies the difference between two group means in terms of standard deviation. Cohen suggested using the following standard deviation values as benchmarks for interpreting the significance of effect sizes:

Table 1 Cohen's d effects

Cohen d value	Interpretation of Effect
0.2-0.3	Small effect
0.5	Moderate effect
0.8	Large effect

These thresholds provide a general rule of thumb, but the interpretation of effect sizes should also consider the context, study design, and field of research. In some disciplines, even small effect sizes may be meaningful (eg medicine), while in others, larger effects are necessary for practical relevance. In addition, average effect sizes in policy domains where we have a large number of trials is substantially smaller than those predicted by Cohen – and so powering to detect small effect sizes on this scale is probably sensible.

While you could just choose which effect size you think your intervention would have, it is sensible to run **repeated power analyses, assuming different effect sizes**. By generating a few different sample sizes, we can quickly ascertain which ones are feasible. Perhaps a trial with a small Cohen's d estimate is unrealistic because of the large required sample size, so if we suspect a small effect, perhaps another evaluation method must be pursued.

Finally, it's worth stepping back and critically appraising the evidence. There is evidence that researchers are often overly optimistic in estimating effect sizes (Sanders et al., 2020). This optimism might be driven in part by the researchers themselves (they may strongly believe the intervention works and estimate accordingly) and by the systematic bias proceeding from academic publishing, which tends to favour studies with statistically significant results over null results. Therefore, as researchers we should proceed with **caution**, and tend towards **less optimism** in estimating anticipated effect sizes.

Introducing complexity: covariates, cluster randomisation, stratification, and attrition

Thus far, we have assumed a straightforward trial, where each unit of analysis is independent of others in terms of their treatment status and any correlating variables. There are many ways to muddy the power analysis picture, however, including cluster randomisation of treatment, stepped-wedge designs, and complex interventions. Here we will focus on including covariates in analysis, clustered randomisation, stratification, and attrition, as the first two are common design elements in trials and the third is a common reality researchers must account for.

Accounting for covariates

Including covariates in your analytical model is common practice in randomised trials. Opinions differ about how many should be included, but in general, most guidance recommends controlling where possible for baseline levels of the outcome measure (collected prior to randomisation and treatment) and controlling for any variables used in stratification. In regression analysis, these variables, included on your analysis, will explain some of the variance in the model, reducing the unexplained variance, and decreasing the minimum detectable effect size.

The extent to which this will happen depends primarily on the proportion of the variance that is explained by the variables included in your model. Controlling for the variables we've described – baseline levels of the outcome measure and stratification variables – will typically do a fairly good job of explaining variance. In the case of baseline measures of the outcome measure, this is because the best predictor of the future is often the past – the people who do best on a maths test in September, are also likely to be the ones that do best on a maths test in the subsequent January. In the case of stratification variables, this is because you should choose your stratification variables based on what it is most important for your sample to be balanced on – that is, those things that, if unbalanced, are most likely to lead to bias in your estimated treatment effect – they should therefore be the most predictive variables of the outcome that you have available at the point of randomisation, and hence are also likely to explain quite a bit of the variance.

Covariates included in the analytical model at the end of the trial explain variation in the outcome measure. For example, a lot of variation in height can be explained by gender, and so including gender in an analytical model reduces the amount of unexplained variance in that model. Reducing the unexplained variance increases the amount of statistical power we have, all else being equal. As such, although we should be parsimonious in our analytical strategy, evaluators should include variables that are likely to explain the most variance in the outcomes. Often these will include baseline levels of the outcome measure – which are often the best predictors – as well as demographic characteristics.

The effect of including covariates on power depends on the correlation between the covariates included and the outcome measure. This correlation is denoted R, and we can calculate the revised MDES using the following equation.

$$MDES_{New} = MDES_{without\ covariates} \cdot \sqrt{(1 - R^2)}$$

As we can see from this equation, the larger R is, the smaller the revised MDES is relative to the original MDES.

Cluster randomisation

In our SMS treatment example above, we had individual families being sorted into treatment and control groups, without much concern for spillover effects. This is an example of an individually randomised control trial (IRCT), where it is practical to allocate treatment on an individual level without fear of contaminating the control group and threatening the validity of the study. In many cases, this type of randomisation will not be possible or desirable. Consider a literacy intervention that takes the form of a taught class, targeting homeless sheltered adults. In this case, the treatment allocation is at a group level (the class), and within this group, there will be correlations of potential outcomes among individuals, driven by local geographic factors such as the availability of other services, urbanisation, crime rate, demographics, etc. This within-group non-independence of outcomes impacts the power of our study, with higher correlations requiring a larger sample size to achieve the same statistical power.

Intra-cluster correlation rate

When we know that there are likely correlations of outcomes between units within a cluster, we refer to this as the intra-cluster correlation, or the ICC. This ICC has a value that can be known, with some digging and use of empirical evidence, but many researchers (understandably) fall back on rules of thumb or educated guesses (Sanders & Vallis, 2023). Yet estimating the ICCR accurately has big implications for a cluster-randomised study: estimate too low, and the estimated sample size risks being too small, but estimate too high, and the cost of running the study could quickly balloon out of control. To understand how, let's take a look at the ICCR and a related quantity – the design effect – in more detail.

In technical terms, the ICCR is a ratio of the relative magnitude of within- and between-cluster variances of the outcome of interest; in other words, it assigns a quantity that describes how similar individuals within a cluster are, relative to how similar individuals are within the entire sample. It will always be expressed as a value between 0 and 1, with 0 meaning no correlation of outcomes. The closer to 0, the better from a sampling point of view; it means we will get more information from each unit, which means a more efficient trial. In a medical context, we might see ICCRs around a 0.02, while in a social science context – such as classrooms and schools within a local authority – the ICCR could be as high as 0.25 or higher.

Similarly to calculating statistical power, we cannot know for certain how outcomes will correlate within and between clusters before running the trial. We can, however, make an educated guess based on the context and drawing upon existing evidence. There are three contributing factors to ICCR:

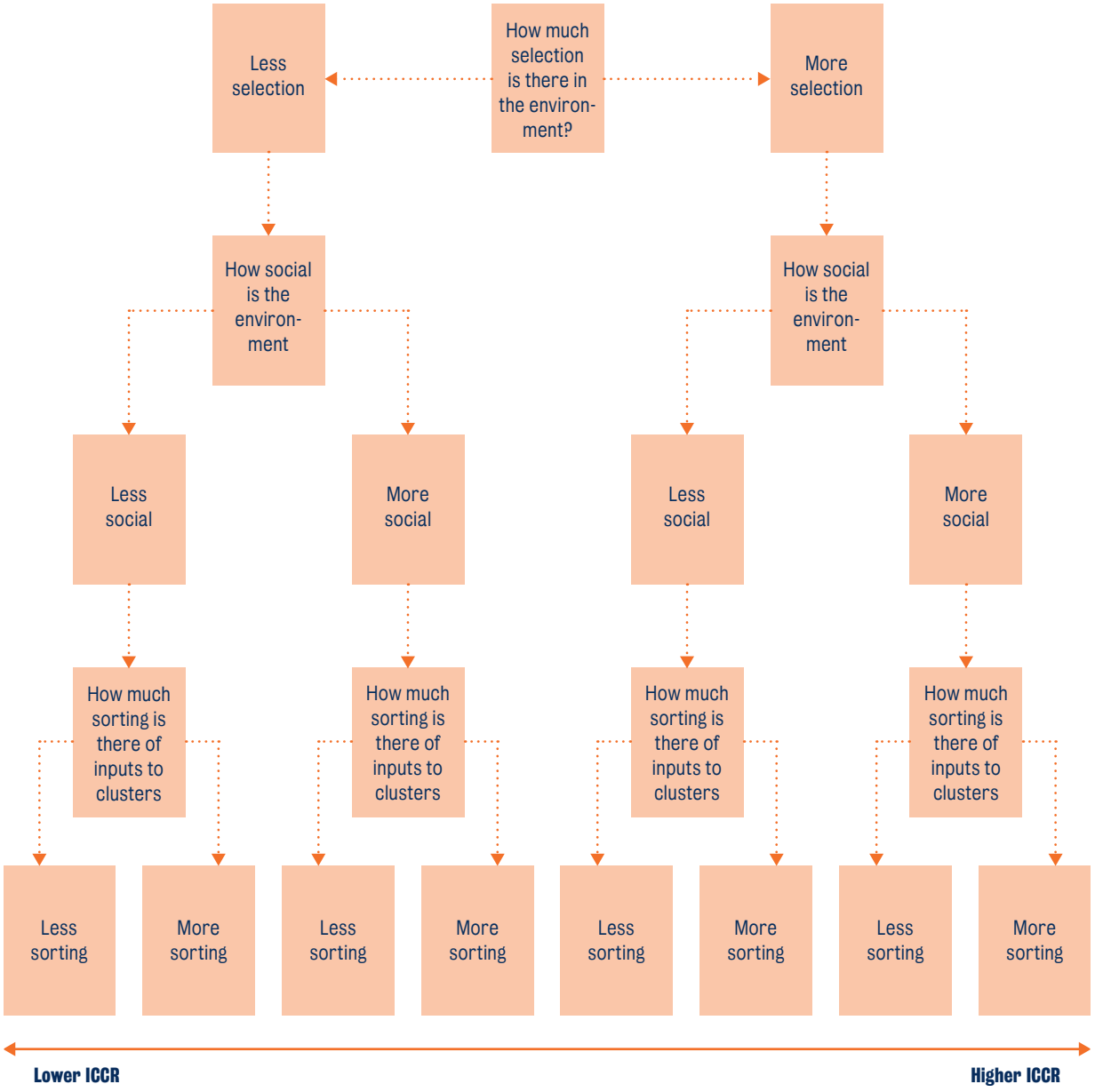
1. The extent to which there is sorting into clusters by participants. For example, if particular types of people tend to cluster together. For example, in schooling, if higher income families tend to send their children to some schools, and lower income families send their children to others; or where there is academic selection, with higher ability children attending some schools and lower ability children attending others.

2. The extent to which there is socialisation of outcomes within a cluster. This relates to what happens when participants are *in* their clusters. For example, anything that happens in a classroom might be highly socialised – with children learning from each-other, or being disrupted by each other. Something like homework – which is done outside of class time, will tend to be less socialised.

3. The extent to which there is selection of the inputs into clusters (teachers into schools, managers into teams, etc.). If good doctors all go to particular hospitals, then the inputs that patients experience will be exhibit more clustering. By contrast, if doctors were randomly assigned to hospitals, then there will be less clustering.

The flow chart on the following page helps to visualise this:

Figure 1 Intracluster correlation rates by degrees of sorting and socialisation



Now that we have a basic intuition of ICCR as a quantity, let's introduce the design effect, which is the ratio of the sample size we would need if we were undertaking an individually randomised trial (Kerry & Bland, 1998). This is the bridge between the ICCR and its implications for the minimum required sample size of the trial. The design effect takes into account the correlation and the number of units within a cluster, and tells us how much larger our sample (which we would already have calculated through power analysis) will need to be based on those values. The design effect can be formulated as:

$$D = 1+(M-1)\rho$$

Where D is the design effect, M is the number of observations in the average cluster, and is the ICC. So, to find the required sample size for a cluster randomised trial, just multiply D against your required sample size if you were running an individually randomised trial.

Stratification and oversampling

In order to ensure that there is balance of certain key characteristics between treatment and control groups within a trial, researchers will often incorporate stratification into their randomisation protocol. Briefly, stratification works by first splitting the sample into groups based on one or more variables, and then randomises treatment within those groups. For example, let's say a researcher wants to ensure treatment allocation across a geographic spread, so they first divide up the sample into counties, and then randomise treatment within each county.

Similar to cluster randomisation, this potentially creates groups within the sample that have similar characteristics, but unlike cluster randomisation, there is a mix of treatment and control units within the group. On its own, stratification may modestly improve power (depending on how you define power, see footnote 1) in small trials in cases where the stratification variable exerts a large influence on the outcome and is included as a control variable in the analysis (Kernan et al., 1999). But there is a limit: over-stratification can create very small subgroups. A smaller n also means fewer degrees of freedom, which will eventually produce a reduction in power as the number of values that are free to vary goes down (ie there is less of a signal to detect within each subgroup).

Unequal allocation

So far, we have mostly assumed that participants are allocated equally to different treatment conditions. In a two-arm trial – with one treatment group and a control group – this means allocating participants at a rate of 50:50 – for every person treated, one person will be in the control group. This is the simplest way of randomising, and also the most statistically efficient – because what matters in determining power is the number of people in each condition, not the sample size overall. Because, as we have said previously, each observation adds less power than the one before, moving sample from one condition to another (for example increasing the treatment group size at the expense of the control group), you lose more power than you gain, and the effect size you're powered to detect increases.

There are, of course, times when uneven allocation is inevitable. This occurs primarily for two reasons. First, when the treatment is very expensive, and budgets are fixed. For example, a programme of intensive therapeutic support for young people might cost £10,000 per participant, and the budget available might be £1 million. In this situation, only 100 participants can be treated in total, but it may be possible to recruit a sample of more than 200. When this is the case, the statistically efficient approach is to maximise the overall sample and unevenly randomise, creating a larger control group. This is because in this instance equal allocation is not a possibility while retaining the larger sample, and the larger sample will deliver more power than a small one with equal allocation.

Second, unequal allocation might be preferred by delivery organisations and intervention developers who believe strongly that their intervention is beneficial, and hence that withholding it is unethical. They will therefore prefer randomisation that favours the treatment group – with more participants allocated to treatment than control. The specific arguments for this will vary from case to case, and are a matter for an ethics committee – but in general it should be avoided. Unless very strong evidence already exists of the intervention's impact, then the ethical argument for uneven allocation is not a strong one, while the ethical argument against running a potentially under-powered trial is clearly relevant.

Attrition

Finally, after conducting a few power analyses with different effect sizes and considering how intra-cluster correlation and stratification may impact power, it is important to consider attrition. If the trial takes place over a period of time and baseline and endline measures are taken, there is often the risk of participants dropping out mid-trial. If the planned sample size does not take this potential attrition into account, it could risk the entire trial becoming underpowered. Similar to estimating the intervention's effect size, researchers can look to other studies on similar populations in order to predict attrition levels and plan accordingly.

Non-compliance

Almost all trials will have some measure of non-compliance, however we choose to think about it. For example, a classroom intervention might have non-compliance in a number of ways. Students participating in the trial might be absent due to illness during some classes; teachers may deviate from the intervention or might not deliver it at all; or students might move from a treated classroom to a control one after randomisation takes place. All of this means that some people who are analysed as though they were in the treatment group didn't receive the intervention, or only received a part of it. Some people may have been assigned to the control group and ended up being treated. While the main analysis of a trial should be conducted on an intention to treat basis (whereby people are analysed based on the condition to which they are assigned, rather than whether they ended up getting treated), it is helpful to think about compliance as well.

To think about how this is going to effect your sample size requirements, we can consider its impact on the effectiveness of the intervention. Let us say, for example, that one in 10 people assigned to the treatment do not comply, and receive 0 per cent of the treatment. Those people will therefore receive 0 per cent of the benefit of the treatment. Our analyses of the impact of the intervention will effectively calculate the average difference between the outcomes of the treatment group and control groups – the average treatment effect. Let's say that the intervention, if received, has an effect of size X. The average treatment effect in the sample we've described – with 10% non-compliance — is therefore 0.9X. .

How this effects the sample size you should aim for in the trial depends on your objectives. If you hypothesise that the intervention might have an effect (if received) of, say, Cohen's d of 0.2, and you want to be powered to detect *this*, you'll need to power your study to detect effects of 0.18 in this instance – to account for the watering down of the effect due to non-compliance – in the case of an individually randomised trial, increasing the total sample size from 786 to 970. If, however, you are basing your effect and sample sizes on the cost of the intervention, and you need to pay for everyone to receive the intervention even if 10 per cent don't receive it, then you shouldn't adjust your sample size to account for this – because although the effect will be watered down, the cost will not be.

Adjustments for multiple comparisons

As we discussed in Chapter 5 when we covered p-hacking the more statistical tests we run, the greater the likelihood there is that we will find a false positive. This is because the 95 per cent confidence interval that we use indicates that there is only a five per cent chance of the positive finding having occurred by chance. However, if we run two tests, then the likelihood of at least one positive result occurring is higher – almost 10 per cent – and this continues to rise with each additional test we run. For example, if we have four main outcomes for a trial, we are running four tests, and so the likelihood of one of them yielding a false positive is higher than if only one test was conducted. To adjust our analysis, and hence our power calculations, for this possibility, the simplest approach is to make use of a Bonferroni correction. To do this, we divide the type one error rate (alpha) by the number of tests that we are running. So, if we take the normal tolerable type one error rate of 0.05, and we are running two tests, we must power our study to detect effects with a type one error rate of 0.025 (2.5 per cent). If we are running four tests, this becomes 0.0125. Other, more sophisticated corrections are available, but using a Bonferroni correction when calculating statistical power is a conservative approach to ensuring well powered studies.

Conclusion

We have made it! We have covered the concept of statistical power, its importance to developing evidence and avoiding type II errors, the basic steps and considerations for conducting a power analysis, as well as some potential threats to a well-powered study. We have developed familiarity with how power interacts with different trial elements, such as larger sample sizes and anticipated effect sizes increasing power, and vice versa for smaller sample sizes and effect sizes. We have also explored how design choices of a trial protocol such as clustered randomisation and stratification can impact power.

Taking a step back for a moment, it is worth reflecting that while technical skills and statistical understanding are fundamental for conducting power analyses, equally important is approaching the process with humility and caution: don't be overly optimistic with anticipated effect size, plan for attrition, and run multiple analyses with different specifications. Finally, take your time: a properly thought-through power analysis contributes to maximising trial resources, respects the time and effort of trial participants and the research team, and helps make the case that a trial's findings are valid.

References

Champely, S. (2020). pwr: Basic Functions for Power Analysis. R package version 1.3-0, https://CRAN.R-project.org/package=pwr.

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences (Second ed.). New Jersey: Lawrence Erlbaum Associates.

Hattie, J. (2008). Visible learning: A synthesis of over 800 meta analyses related to achievement. New York: Routledge.

Kernan, W.N., Viscoli, C.M., Makuch, R.W., Brass, L.M., and Horwitz, R.I. (1999). Stratified Randomization for Clinical Trials. Journal of Clinical Epidemiology: 52(1), p. 19-26. https://doi.org/10.1016/S0895-4356(98)00138-3.

Kerry, S.M. and Bland, J.M. (1998) Statistics Notes: Sample Size in Cluster Randomisation, BMJ: 316(7130), p.549.

Lipsey MW, Wilson DB. (1993). The efficacy of psychological, educational, and behavioral treatment. Confirmation from meta-analysis. Am Psychol: 48(12):1181-209. doi: 10.1037//0003-066x.48.12.1181.

Robinson D, Hayes A, Couch S (2023). broom: Convert Statistical Objects into Tidy Tibbles. R package version 1.0.4. https://CRAN.R-project.org/package=broom.

Sanders, M., Mitchell, C., & Aisling, NC. (2020). Effect Sizes in Education Trials in England. Available at SSRN: https://ssrn.com/abstract=3532325 or http://dx.doi.org/10.2139/ssrn.3532325

Sanders, M. & Vallis, D. (2023). 'Intra-Cluster Correlation Estimates for the Statistical Design of Trials in Household Homelessness.' European Journal of Homelessness Volume: 17(3). https://www.feantsaresearch.org/public/user/Observatory/2023/EJH_17-3/EJH_17-3_RN01_v0163.pdf

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). 'Welcome to the tidyverse.' Journal of Open Source Software: 4(43), 1686. doi:10.21105/joss.01686.

Wiliam, D. (2008). 'International comparisons and sensitivity to instruction', Assessment in Education: Principles, Policy and Practice: 15(3), 253–257.

Chapter 7. Missing data

Introduction

Missing data in randomised controlled trials (RCTs) is a critical and persistent challenge that, if not properly addressed, can significantly undermine the validity, reliability, and interpretability of study findings. The presence of missing data is a common issue that researchers frequently encounter, and it requires careful consideration and strategic handling to prevent it from compromising the overall quality of the research.

It is important that researchers designing trials include a plan for managing missing data, as (a) some level of missingness is all but inevitable in any trial that involves primary data collection, and (b) the decision of how to handle missingness creates a risk of a 'garden of forking paths', whereby researchers can make decisions at the time of analysis (that is, once they have the data), which allow for bias to creep into findings.

When faced with missing data, a researcher's initial instinct may often be to simply remove any cases with incomplete information through listwise deletion. However, while this approach might seem straightforward, it is not always the most appropriate or effective solution. Listwise deletion involves excluding all data from any participant who has a missing value for any variable involved in the analysis. This method, although the most straightforward, implies the loss of potentially valuable information, which can result in decreased statistical power and in some cases biased estimates, where those who are missing are not the same as those who are not, in a way which influences the finding of the study. Thus, deletion of cases with missing data, also known as a 'complete case analysis', can cause more harm than the missing data itself by introducing further biases and inaccuracies into the results.

Much like other crucial steps in the design and analysis phases of a trial, addressing missing data requires a methodical approach. Researchers must follow a series of logical steps to properly manage missing data, starting with identifying, as far as possible, the 'type' of missingness. Understanding whether data are missing as the result of chance or if missingness is conditional on observed or unobserved covariates is essential in determining the most suitable method for handling missing data. Consequently, the ability to identify the extent of the problem they are faced with is a critical skill that trialists must develop to ensure that any adjustments made to the data do not inadvertently introduce bias or reduce the precision of the study's findings.

Moreover, an in-depth understanding of missing data mechanisms can greatly influence the outcome of a trial by helping researchers implement strategies that effectively mitigate the risks associated with them. Without this knowledge, there is a heightened risk that the study's results will be compromised, either through the introduction of confounding or through a loss of precision in the estimates. Therefore, the ability to address missing data appropriately is not just a technical skill but a fundamental aspect of ensuring the integrity and success of a trial.

Understanding missing data mechanisms

Each type of missingness is the consequence of different patterns within the data and as a result has different implications for a study. This implies each type requires a different approach to mitigate its impact on the study's outcomes. There are usually three types of missingness encountered in studies, missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR), however in randomised controlled trials, there is also a sub-type of missingness within MNAR that may be encountered, missing experimentally not at random (MENAR). We will now proceed to cover each of these in turn.

Missing completely at random (MCAR)

MCAR occurs when the probability of missing data is independent of both observed and unobserved data. This type of missingness is purely random and does not correlate with any participant characteristics, outcomes, or treatment assignments. For example, if data from a random set of participants is lost due to a computer error that is unrelated to their treatment or outcomes, the data is MCAR.

For data that is MCAR, analysis based on complete cases (participants for whom all data points are available) will be unbiased, however listwise deletion will also reduce sample size and hence statistical power.

The extent of the reduction in power will depend on the type of study (ie individual-level vs. cluster-level randomised trial) and the size of the missing sample. Due to the random nature of the missingness, no imputation is needed if the data is MCAR as long as the size of the missing sample is not large. A general rule of thumb suggests that listwise deletion may be appropriate if less than 5 per cent of the data is missing. In that case, analysis can proceed with complete case data, though researchers may opt for imputation techniques to maintain power. Alternatively, models such as Full Information Maximum Likelihood estimation can allow for the full use of all available data, bypassing the need to exclude available information as a result of listwise deletion in a wide format; it is important to note that this method does not impute any data, but rather uses each case's available data to compute maximum likelihood estimates. Alternatively, imputation of baseline outcome scores and/or demographic characteristics will not introduce bias into the estimate of a treatment effect if data are MCAR, and so, imputation approaches such as multiple imputation by chained equations (MICE) (Rubin, 1987), can be safely used.

Determining if the data is MCAR relies not only on conducting a series of statistical tests attempting to exclude the likelihood of other types of missingness, but also considering all available information with respect to the conducting of the trial itself. If researchers are aware of the underlying reasons for the missingness (as in the example provided above) and can confirm they are independent of other covariates, then moving forward with the assumption of MCAR may be acceptable.

Missing at random (MAR)

MAR occurs when the probability of missing data is related to observed data, but not to unobserved data. In other words, the missingness is associated with known characteristics, but it is conditionally independent of the unobserved data once the observed data is accounted for.

For example, if participants with lower baseline health scores are more likely to miss follow-up assessments, but these baseline scores are recorded, the data is MAR – because the missingness is explained by the baseline scores. In the case of MAR data, complete case analysis may in some cases lead to unbiased estimates, but the loss of power is more significant. This is because in the case of MCAR data, missingness is completely random which implies the remaining data after excluding incomplete cases is still representative of the overall population and the correlation structure of the covariate(s) where the outcome is maintained, a fact which is not true in the case of MAR data whose missingness is conditional on observed information. For example, suppose we are investigating the effect of an intervention on health status and include a covariate of self-reported income in our model at baseline to improve model accuracy. If a subgroup of participants identified by covariate Z are less likely to disclose higher levels of income, then failing to account for Income missingness due to Z will distort the correlation between income and health status. This effect implies that, combined with the reduction in power due to a smaller sample size (as in MCAR), a greater part of the outcome's variance now remains unexplained leading to further power loss.

Just as importantly, if the characteristic correlated with missingness is also correlated with effect size, then this may influence the size of effect estimated, and hence alter the external validity of the study. To maintain the integrity of the analysis and guard against potential introduction of bias, imputation methods should be employed by multiply imputing or null imputing missing numerical covariates using MICE. In many cases, this may include covariates which would not be used in a particular analysis. For example, if a study involves the analysis of three outcomes using models 1 and 2, and some demographic covariates are intended to be utilised as part of the analysis only for model 1, then these additional covariates can also be included as part of imputation for model 2, and vice versa; these covariates are often called auxiliary variables as they can be part of the imputation without being part of the model's analytical specification. It is important to clarify that the goal of imputation is not causal inference but accurate prediction, and as a result, it is acceptable to use any inputs in the imputation model to achieve this goal.

Multiple imputation (MI) or other appropriate imputation techniques can be used to handle MAR data. These methods help maintain the sample size and preserve statistical power by imputing missing values based on observed data. Identification of which observed covariates are predictive of missingness can be achieved by constructing an outcome variable coded as 1 if data are missing, 0 if otherwise, and then regressing on observed covariates through a logistic regression. If statistical significance is found for the estimated coefficient(s) on the likelihood of missingness, this suggests the covariate(s) used in the regression can –to some extent- predict missingness and then data is likely MAR or a mixture of MAR and MNAR.

Missing not at random (MNAR)

MNAR occurs when the probability of missing data is related to unobserved data, making it the most challenging type of missingness to handle. The missing data mechanism is linked to factors that are not observed, which may lead to significant bias if not appropriately addressed. For example, if participants who experience worse outcomes (unobserved) are more likely to drop out of the study, and this is not associated with any recorded baseline variables, the data is MNAR.

MNAR data can introduce bias into the analysis, as the missing data is systematically different from the observed data. Sensitivity analyses are crucial to assess the robustness of the study's findings. Techniques such as pattern-mixture models or selection models, which make assumptions about the missing data mechanism can be used. These methods often require external data or expert judgement to validate the assumptions.

In real-world situations, researchers may be faced with a mixture of MAR and MNAR. In this scenario, simulation studies have shown that even if some data are MNAR, multiple imputation may still aid with mitigation of bias and improvement of precision. Collins, Schafer, and Kam (2001) explored the issue of MNAR in the context of linear regression. Their findings indicated that when the missing data rate was 25 per cent or less and the correlation between the unobserved covariate and the outcome was 0.4, the imputation model had a negligible impact with regression coefficients and variance estimates remaining unaffected. In more extreme cases, however, for example when the missing data rate reached 50 per cent or the correlation was as high as 0.9, the effects were more substantial.

The best course of action, therefore, is to always firstly assume the data is likely MAR and test for that assumption by regressing a binary missingness identifier with other covariates. Assuming some significance is found, proceed with MI and compare resulting estimates with those of complete case analysis to check for robustness. If results from the imputed dataset deviate, then a combination of MAR and MNAR is the more likely scenario. If MNAR is suspected to be the more dominant mechanism, one could aggregate data by including cluster means for missing baseline covariates and conduct sensitivity analyses to gauge robustness. Alternatively, simulations customised for the design of interest can also be conducted, (see Collins, Schafer, and Kam (2001) to test at which level of correlation and missingness are the parameters of interest affected.

Missing experimentally not at random (MENAR)

MENAR occurs when the missingness of data is directly related to the treatment effect or the treatment assignment itself. This form of missingness poses a significant challenge as it can introduce substantial bias into the study's results, particularly if there is differential attrition between the trial arms. Differential attrition occurs when participants in one group, often the control group, are more likely to drop out or miss follow-up assessments than those in the treatment group, potentially due to the lack of treatment itself.

For instance, consider a scenario where participants in the treatment group who experience adverse effects or discomfort are more likely to skip follow-up assessments or drop out of the study altogether. In such cases, the data is missing in a manner that is not random but rather systematically related to the treatment. This scenario is a commonly encountered example of MENAR, where the missing data reflects a specific pattern influenced by the treatment's impact on participants.

The presence of MENAR can severely bias the estimates of the treatment effect if not properly addressed. This bias arises because the missing data may disproportionately represent certain outcomes, leading to over or underestimation the treatment's true effect. For example, if adverse effects lead to higher dropout rates in the treatment group, the remaining data may falsely suggest that the treatment is more effective than it actually is, as the most adversely affected participants are no longer represented in the analysis.

To mitigate the bias introduced by MENAR, one approach that researchers can employ is the imputation of missing data within the treatment groups. This method involves imputing or estimating the missing outcomes based on the observed data from other participants within the same treatment group. By doing so, the imputation model accounts for the treatment effect when generating the missing values, helping to reduce the bias that could otherwise distort the study's findings without injecting information from other arms, whose outcome is not conditional to treatment. This approach is particularly valuable in maintaining the integrity of the treatment effect estimates, as it allows for a more accurate and fair comparison between treatment and control groups. Once imputation has taken place, compare the treatment estimates using the imputed data and data restricted to the complete cases. If the results are similar (ie significant / insignificant, the same direction and only differing in magnitude of up to 20 per cent), then data is likely to be missing at random (MAR). If the results are dissimilar, the data is likely to be missing not at random (MNAR).

Multiple imputation by chained equations (MICE)

There are a number of multiple imputation techniques that can be utilised, traditional MI methods often rely on the assumption of a joint normal distribution, which may not work well for large datasets with diverse variable types. MICE offers a flexible and powerful alternative that is suitable for datasets with many missing observations and various variable types. The strength of MICE draws from its use of a series of regression models, where each variable with missing data is modelled based on the other covariates that may or may not be part of the main analysis model (auxiliary). MICE utilises an iterative process or a 'chain' of regressions that cycles through all covariates requiring imputation while utilising an appropriate model in each case depending on the type of variable imputed (ie logistic regression for binary variables). This process is repeated N number of times after which an imputed dataset is generated. Researchers then set this process with the goal of generating some number of imputed datasets M (usually for M=10). Main model analysis is then run using each of those datasets to generate M model estimates, after which using Rubin's Rules (Rubin, 1987) they are pooled in order to generate the final model estimates.

MICE follows a step-by-step process broken down into four general steps:

- Step 1: A simple imputation, such as imputing the mean, is performed for every missing value in the dataset. These mean imputations can be thought of as intermediate or 'place holders'.

- Step 2: The 'place holder' values from the variable X in Step 1 are set to missing and regressed on the other variables in the imputation model, which may or may not consist of all variables in the dataset. These regression models operate under the same assumptions that one would make when performing linear, logistic, or poisson regression models outside of the context of imputing missing data.

- Step 3: The missing values for X are then replaced with imputations from the regression model. When X is subsequently used as an independent variable in the regression models for other variables, both the observed and these imputed values will be used.

- Step 4: Steps 2–3 are then repeated for each variable that has missing data. The cycling through each of the variables constitutes one iteration or 'cycle'. At the end of one cycle

all the missing values have been replaced with predictions from regressions that reflect the relationships observed in the data.

- Step 5: Steps 2–4 are repeated for a number of cycles, with the imputations being updated at each cycle, generating an imputed dataset.

The entire process of steps 2-5 is repeated to generate an M number of imputed datasets. The main model analysis is then run on each of these M datasets and resulting estimates are pooled to generate the final estimates.

Addressing missing data in RCTs

The approach to handling missing data varies depending on whether the missingness affects outcomes, baseline or demographic covariates. Each type of variable serves a particular purpose to the study and the reasons for its inclusion play an important part in determining how best to handle its missing values.

Outcome missingness

Outcome data missingness is one of the most challenging types of missingness as it may directly impact the ability to assess treatment efficacy. The appropriate handling of missing outcomes is critical to avoid bias and ensure valid conclusions. The following steps provide a framework for assessing the extent of the problem and mitigating any potential confounding successfully. Additionally, it is important to note that for outcome data missingness, imputation aims to address bias but not power.

Step 1) Assess the missingness mechanism:

Regress an indicator of missingness on baseline covariates utilising a logistic regression.

- If any of the coefficients on the covariates of the missingness model are statistically significant at the 5 per cent level, then data is likely not missing completely at random (MCAR).

- If the treatment dummy is significant, consider MENAR, which requires imputation within treatment groups.

Step 2) Imputation strategies:

- For MCAR: Listwise deletion can be used, as it does not introduce bias, but depending on the extent of missingness, it may affect power. Multiple imputation techniques can still be used if missingness is high.

- For MAR: Use multiple imputation, where missing outcomes are estimated based on other observed data. It is also acceptable to exclude the missing cases, as long as all the variables that affect the probability of missingness are accounted for in the regression.

- For MENAR: Impute within treatment groups, ensuring that the imputation model accounts for treatment effects.

Step 3) Sensitivity analysis:

- Conduct sensitivity analyses to assess the impact of different imputation strategies on the study's findings.
- Techniques like baseline observation carried forward (BOCF) can be used where the missing outcome value is replaced with the observation at baseline
- Alternatively, control drifted observation carried forward (CDOCF) can also be used. This is done by replacing the missing outcome with the baseline observation and then adding the 'drift' from the comparator group to take into account the amount their outcome would have changed in the absence of treatment. The drift is calculated using an autoregression model for the comparator group.

Baseline covariate missingness

Baseline covariates are essential for adjusting treatment effect estimates and for stratification in RCTs. Missing baseline data can affect the balance between treatment groups and reduce the precision of estimates (loss of power).

Step 1) Assess the missingness mechanism:

- Regress an indicator of missingness on other covariates and outcomes.
- If any coefficients are statistically significant, the data is not MCAR.

Step 2) Imputation or aggregation:

- If the covariate is MAR, use multiple imputation and compare results against listwise deletion; if there are no substantial changes, the data likely MNAR.
- If the covariate has substantial missingness (>90 per cent), consider dropping it if pre-specified in the trial protocol.
- For MNAR, use sensitivity analysis and potentially aggregate data by including cluster means for missing baseline covariates.

Step 3) Power considerations:

- If the data is MCAR or MAR, complete case analysis is likely unbiased but less well-powered.
- Use a missing category for categorical variables and multiply impute or null impute missing numerical covariates to maintain power. Where aggregation can be used to preserve power, in general it should be. An example of this might be with missing baseline data including a covariate that takes the value of the baseline measure for those for whom it is observed, and the cluster mean value for those where it is not.

- Whether clustering should be taken into account when imputing depends on a number of factors such as cluster size, variability of follow-up rates between clusters and the intra-cluster correlation estimate (Taljaard, Donner and Klar, 2008).

Demographic data missingness

Demographic data are often used for subgroup analyses and ensuring generalisability of the trial results. Missing demographic data can lead to concerns about the representativeness of the sample.

Step 1) Assess the missingness mechanism:

- Similar to baseline covariates, regress an indicator of missingness on other covariates and outcomes to determine if the data is MCAR, MAR, or MNAR.

Step 2) Imputation or categorisation:

- For MAR, use multiple imputation. For categorical variables, create a missing category if the data is MAR or MCAR.

- For MNAR, sensitivity analysis is crucial to understand how missing demographic data might bias subgroup analyses or generalisability.

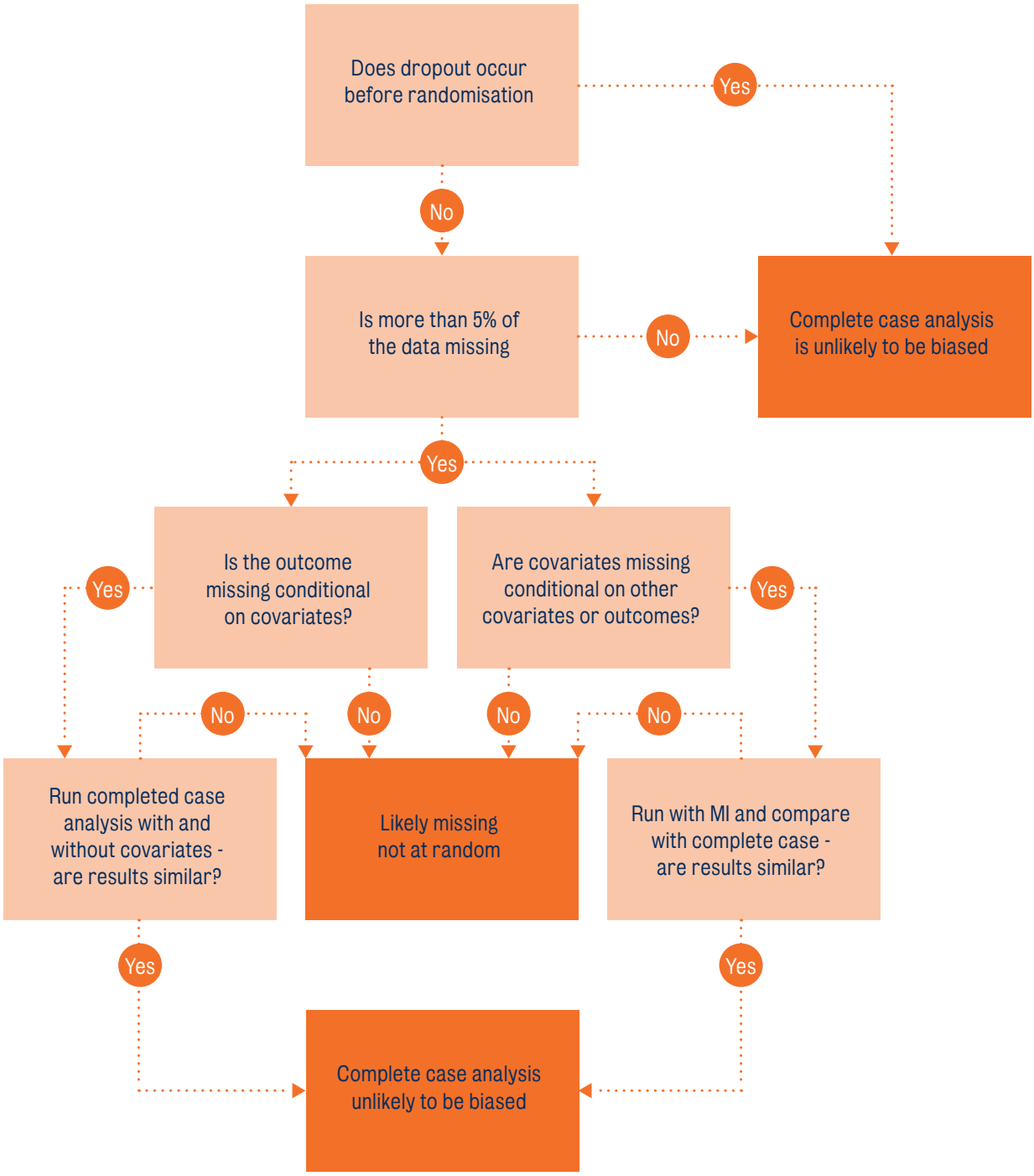
Step 3) Reporting:

- Report the extent of missing demographic data and the method used to address it. Transparency is essential for assessing the potential impact of missing data on the study's findings.

Flowchart for addressing missing data in RCTs

Below is a flowchart that outlines the decision-making process for handling missing data in RCTs.

Figure 1 Handling missing data in RCTs



Reporting

In addition to the analysis discussed, it is beneficial to provide summary statistics that enable a qualitative assessment of the impact of missing data on the treatment effect estimates.

Specifically:

- For each outcome and covariate, report the percentage of missing data along with relevant summary statistics.
- Report the rates of loss to follow-up or missing 'post-'data collection for each arm of the trial.
- Provide a table of baseline characteristics, broken down by treatment arm, and further categorised by response to 'pre-' and 'post-'data collection. This should include both participants who remained in the study and those who joined, to determine whether the characteristics have become more imbalanced by the time of 'post-'data collection.

Conclusion

Missing data in RCTs presents significant and multifaceted challenges that, if not properly addressed, can severely compromise the study. The impact of missing data extends beyond simple data loss –it can introduce bias, reduce statistical power, and ultimately distort the conclusions drawn from the trial. It is therefore critical when dealing with missing data to thoroughly understand the underlying mechanism behind the missingness, its extent and the types of covariates it is affecting. This understanding is essential as it guides the choice of the most appropriate and effective methods for addressing the missingness.

The strategies for dealing with missing data – whether through complete case analysis, various imputation techniques, or sensitivity analyses – must be carefully selected and implemented to ensure that the study's conclusions remain robust, reliable, and valid, despite the challenges posed by missing data. Researchers should also maintain a high level of transparency regarding the extent of the missing data, the assumptions underlying the chosen methods for handling it, and the potential impacts on the study's findings. By adhering to systematic approaches, rigorously applying the chosen methods, and conducting thorough sensitivity analyses, researchers can effectively mitigate the risks associated with missing data, thereby preserving the integrity of the study's conclusions.

References

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychological Methods, 6(3), 330-351.

Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. New York: Wiley. https://doi.org/10.1002/9780470316696

Taljaard, M., Donner, A., & Klar, N. (2008). Imputation strategies for missing continuous outcomes in cluster randomized trials. Biometrical Journal, 50(3), 329-345. https://doi.org/10.1002/bimj.200710423

Chapter 8. Power and inequality

Introduction

Criticism of the use of RCTs comes from both sides of the quantitative/qualitative divide. Some empirical social scientists, including Deaton and Cartwright (2018), critique RCTs as an empirical approach, particularly in terms of their external validity. On the other side, RCTs have drawn criticism for being reductive, and erasing the lived experience of individuals involved.

While 'flattening' the experience of individuals into a single estimate of effectiveness is in some ways the purpose of an RCT (and much causal analysis), even staunch advocates should be concerned about this when participants who might be especially vulnerable, disadvantaged, or discriminated against might make up a minority of a trial's sample, and consequently their erasure might serve to perpetuate cycles of disadvantage.

Many social scientists are committed to conducting more and better research to understand the differential impact of interventions on people from different ethnicities and races. This is in some ways in tension with the rise of RCTs, where subgroup analyses are typically underpowered, and consequently unable to detect all but very large differences in effectiveness between groups.

This short chapter considers the impact of clustering on the power implications of such subgroup analysis, and makes recommendations for how research in this area might be expanded.

People who are members of different ethnicities experience different outcomes, but this picture is complicated. Some ethnicities are much more likely to experience state intervention in family life than others (Bywaters et al., 2019); some do better in school than their white counterparts, while others do worse (ONS, 2020); rates of poverty differ between groups similarly (ibid). The same is true in health settings (DHSC, 2021).

Less thoroughly tested, but no less valid a hypothesis, is that these groups might differ in their responsiveness to interventions, either at the level of practitioners' and professionals' activities, or at a policy level. To the extent that interventions to improve, for example, educational attainment, replicate (or overturn) the conditions that cause differences in outcomes, these interventions are likely to be differentially effective for different groups of participants. Treating study participants from a wide variety of ethnic groups ('BAME') as a homogeneous mass could therefore be statistical bad practice as well as risky.

This concept is not merely of theoretical or intellectual interest. If interventions that 'work' – as judged by an RCT – overall are ineffective for any group which already experiences worse outcomes, then the effect of these interventions being rolled out more widely will be to prolong and exacerbate existing differences between groups; we 'randomistas' will contribute to the systemic injustice in the world.

This being the case, we have a duty to, alongside other considerations, examine whether there are differential effects between groups, and to tailor our recommendations accordingly. In the next section, we outline the challenge faced in looking at subgroup analyses, and how this grows especially challenging when we want to answer questions about more than two groups. Following this, we consider a special case – the cluster randomised controlled trial – and how this might offer a good starting point for this research.

The problem of causal identification

It is impossible to identify an effect on an individual. When you have a headache, and you take an aspirin, and your headache goes away. Was that caused by the aspirin? Was it caused by the passage of time? Or the glass of water you took with the aspirin? We cannot tell.

If we want to find causality, the individual is not a place we can look. Instead, we must look at a group – a sample. In the simplest form, we give half the sample (chosen at random), an aspirin when they have a headache, and the other half we do not – we give them a placebo, and have them drink the water. These two groups are the same, except the aspirin that one group got, and so we say that any differences between them are caused by the aspirin.

In order to say this, we flatten the characteristics of the individual, through looking at averages, and by assuming away these differences into the moments of a distribution. This is possible because we accept the existence of a distribution. Every person is unique, but when we are joined together, our properties are a part of a great distribution. It is the distribution that we seek to replicate in order to find effects, not the individuals that make it up.

When we want to look at particular groups, we break the distribution down into parts based on some shared characteristic: gender, or race, or baseline scores. Each time we break down the distribution, we have predictable effects, we reduce the flattening of the data, we make the assumption of distributional equality harder to be sure of (meaning that causal inference is less certain) and we reduce our ability to detect small differences.

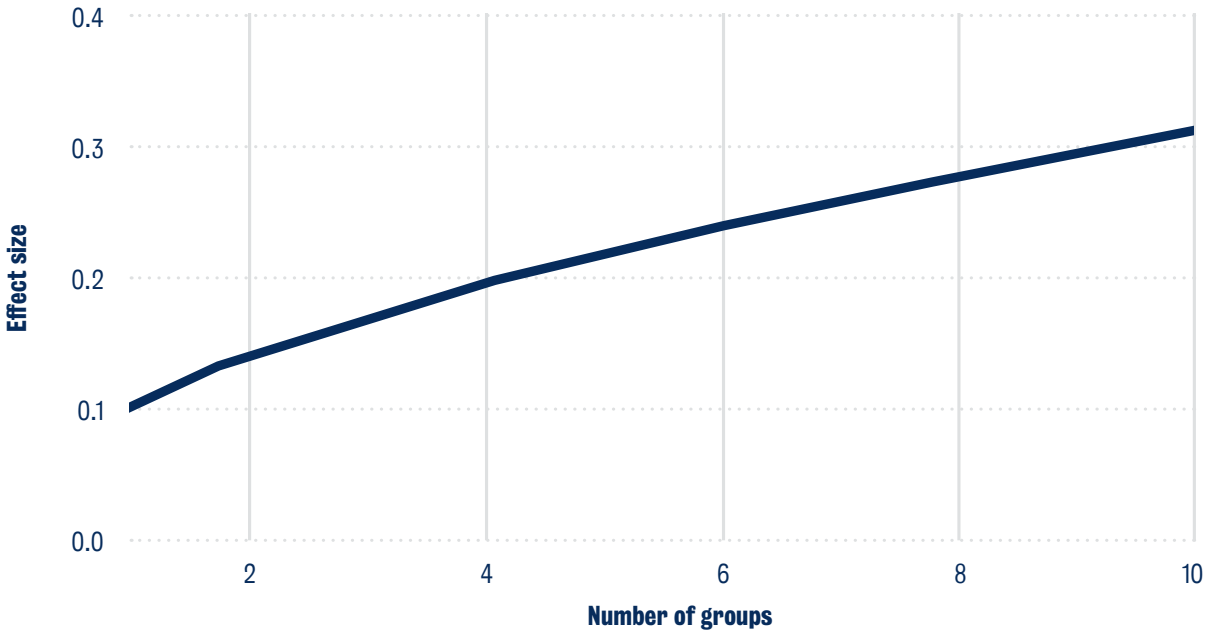
Because of this, we face a trade-off; the more 'representative' of the individuals within the data our analysis is, the harder it is to detect effects, and the harder it is to have confidence in the rigour of our research. The sections below attempt to articulate the trade-offs that we face.

Effect sizes

The more groups that a sample is divided into, the less statistical power we have to detect effects of a given size. This means that, for example, the positive benefit for a group would need to be larger for us to find it as statistically significant. The impact of this is that the smaller the group we break down to, the harder it is for us to say that an intervention is successful, and the harder it is for us to find support for further scaling. Statistical significance is not the only threshold that we can or should use to measure effectiveness, but it is a useful, and consistent metric, and so it is the one we use to illustrate the point here.

The graph below shows the effect size that can be detected in an individually randomised controlled trial, as we divide the group up. As a reference point, we assume that the study is designed to detect an effect of 0.1 standard deviations, which is a minimum detectable effect size (MDES) in line with average effect sizes in education randomised trials (Sanders et al., 2020). As the graph moves to the right, it shows the rise in MDES for each individual group as we divide the sample into more, smaller groups of equal size.

Figure 1 Minimum detectable effect size by number of groups



Clustering and subgroups

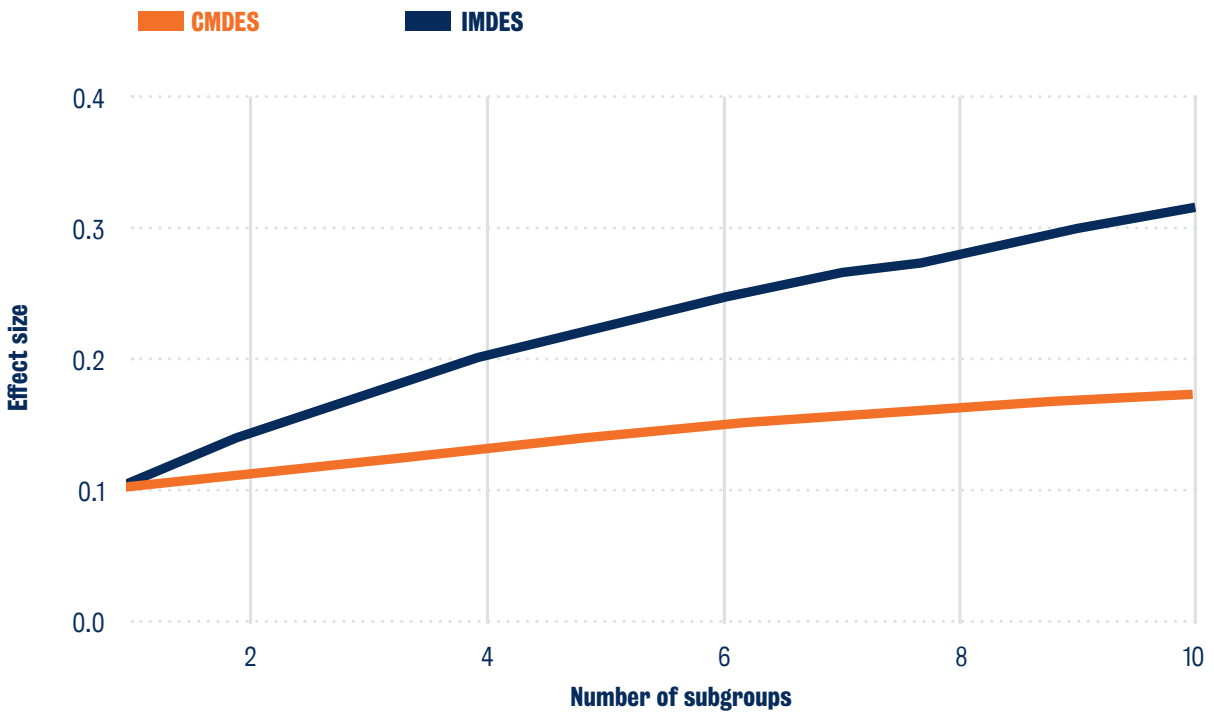
We have seen that statistical power is a challenge when looking at smaller groups, and consequently analysis the smaller groups we divide the sample into, the less rigorous each analysis is. Current and historic studies lack the sample to allow confidence for re-analysis in a lot of cases, and future studies face an inevitable trade-off between cost, representativeness, and being able to conduct analysis of this type.

When we are interested in race, we may find that cluster –randomisation (for example, randomisation at the level of a school, firm, or locality), typically an enemy of well- powered studies, provides a potential route by which current and historic trials might be re-analysed to look at effects for minority groups while simultaneously retaining statistical power (although it should be noted that for trials randomised where race mostly varies across rather than within clusters – for example families – this benefit will not exist).

Let us consider the loss of statistical power from two trials. Both trials are powered to detect effects of 0.1 standard deviations in the main (whole-sample) analysis. One is individually randomised, and the other randomised at the cluster (let's say, the school) level – the same as the one we described above. We can assume, for the sake of argument, that the size of a cluster is 30 participants, and that the intra-cluster-correlation rate (ICCR) is 0.1. For simplicity we follow Kerry and Bland's (1998) simplified form. To achieve this level of power, the cluster randomised trial will be much larger in terms of numbers of participants, but we take this as given for our purposes.

Now, we can consider analysis of subsamples of various sizes, and how power (represented here by minimum detectable effect sizes, MDES) deteriorates in each case, presented in the graph below, which assumes for simplicity even distribution of subgroup members across clusters.

Figure 2 Minimum detectable effect size by number of subgroups (individual and cluster-randomised)



The implications of this graph and its underlying analysis are clear: that if one is conducting or commissioning a cluster randomised controlled trial already, or if analytically, a new trial makes most sense to be a cluster randomised controlled trial, then it simultaneously makes sense for the evaluation to look at subgroup effects, and to potentially expand the evaluation to consider these smaller groups.

Conclusion

This chapter has presented a very brief analysis showing that the trade-off between statistical power and representing the experiences of non-white participants is less stark for cluster randomised controlled trials than it is for individually randomised trials

The analysis presented in this chapter is trivial enough that it is likely to be known to anyone who has ever conducted a power calculation for a field experiment. However, the implications that we present here – that these trials give us a better chance to understand and combat inequalities – are novel.

There is an important limitation to this analysis in that it proposes a very simple analytical specification and relies on two core assumptions: first, that the ICCR remains the same in subgroups as it is in the whole group; and second, that there is even spreading of subgroups throughout clusters. These assumptions make for an analytically straightforward path to the results we present here, but deserve empirical verification.

Either having conducted this empirical verification, or alongside it, this suggests two routes forward. First, it suggests that it could be a fruitful avenue of research to re-analyse existing cluster randomised controlled trials, to look at differential effects by race and to conclude whether these exist, and to look at whether particular interventions work better – or worse – for members of minority ethnicities. Second, it suggests that research designers, and importantly funders, seeking to do more to improve their research's contribution to racial equality, should allow for, and indeed fund, analysis that looks more deeply at race, especially in the context of cluster randomised controlled trials. This would not be a trivial improvement: of 83 studies funded by the UK's Education Endowment Foundation and considered by Sanders et al. (2020), 56 were cluster randomised.

References

Kerry, S. M., & Bland, J. M. (1998). Statistics notes: Sample size in cluster randomisation. BMJ, 316(7130), 549.

Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. Social Science & Medicine, 210, 2-21.

Bywaters, P., Scourfield, J. B., Webb, C., Morris, K., Featherstone, B., Brady, G., ... & Sparks, T. (2019). Paradoxical evidence on ethnic inequities in child welfare: Towards a research agenda. Children and Youth Services Review, 96, 145-154.

ONS (2020): 'Child Poverty and Education Outcomes by Ethnicity'
https://www.ons.gov.uk/economy/nationalaccounts/uksectoraccounts/compendium/economicreview/february2020/childpovertyandeducationoutcomesbyethnicity

Department for Health and Social Care (2021). Ethnicity Facts and Figures Service: Health
https://www.ethnicity-facts-figures.service.gov.uk/health

Sanders, M., Mitchell, C., & Ni Chonaire, A. (2020). Effect Sizes in Education Trials in England. Available at SSRN 3532325.

Chapter 9. Consent, assent and randomised evaluations

Introduction

In this chapter we consider ethics of consent and assent in RCTs. We are particularly and principally concerned with the ethics of (parental) consent and assent for research with children and young people (CYP), but there are broader applications for other vulnerable individuals or people with diminished capacity. This focus reflects two realities – first, that these issues are more prevalant among research with young of vulnerable people, and second that the what works movement thus far has been particularly impactful in areas of policy that relate to young(er) people. Of 13 what works centres currently established in the UK, six are either exclusively or majorly interested in outcomes for people under the age of 25.

This focus on younger people, and indeed on a particularly vulnerable subset of younger people, and the context in which new 'interventions' or 'treatments' take place marks a radical departure in the conduct of randomised controlled trials compared with their dominant use case – the testing of interventions in medicine.

Translating trial methodology from one field to another has not been without challenges. Although these are best recounted elsewhere (see eg Edovald and Nevill, 2021, or Sanders and Breckon, 2023), it is worth illustrating that they are widespread. Trials with young people in social policy settings are larger, some involving tens of thousands of students (eg Miller et al., 2017, Stokes, 2022). Assumptions of effect sizes translated from medical settings led early trials to be underpowered (Sims et al., 2022, Sanders et al. 2020), as effects of diffuse, interpersonal interventions, are smaller than the effects of drugs administered directly to patients in hospitals or primary care settings. Another challenge grappled with by researchers relates to the ethics of running trials in these contexts and with these participants and is the topic of this chapter.

Background to ethics and consent

Since the end of the Second World War, the formalisation of research ethics has led social science research to be conducted in a way that is more standardised and more ethical in the way that it approaches individual participants – not least reflected in the transition from describing those involved in studies as 'subjects' – to whom research happens – to describing them as 'participants' who are an active part of research. Deception, a common feature of early laboratory experiments in psychology, is no longer permissible beyond the life of the study (that is, if people are deceived, they must be debriefed) (Kelman, 2017) and is forbidden in laboratory experiments in economics – people must not be deceived at any point (Bonetti, 1998).

Use of parental opt-out assent alongside opt-in consent from schools and pupils is widespread in educational research, but Institutional Review Boards in the United States and Research Ethics Committees in the UK vary in their comfort level approving studies with this design, and researchers may face longer and more challenging ethics approval processes. As school-based studies often need to start at a specific time (eg the start of the school year), uncertainty in how long ethics approval will take, or whether it will be granted, add considerable complexity to the set-up of these studies. Our goal with the chapter is to consider the circumstances under which opt-out assent from parents is likely to be methodologically necessary and can be ethically permissible – and potentially ethically superior.

The chapter proceeds as follows. After defining opt-in consent and opt-out assent, and considering the prevalence of the two approaches in educational research, we outline the ethical, methodological, and practical considerations that may affect the choice of consent procedure. We argue that although there are substantial ethical challenges to relying on opt-out assent from parents, there is also an ethical case that in certain circumstances it is more appropriate as it is fairer and more inclusive, and ensures that CYP from underrepresented backgrounds are not disproportionately excluded from activities they may wish to participate in, and that undue burden is not placed on schools. We then consider the methodological reasons to prefer opt-out assent from parents, which primarily relate to sample size, representativeness, and practical considerations relating to data protection, budget and logistics. Finally, we discuss how to weigh up these considerations, and how to mitigate ethical risks related to using opt-out assent for parents.

Opt-in consent

Opt-in consent to participation in research requires that a person is informed of the nature of the research that they are taking part in (even if they may not know exactly what treatment they are going to receive). Participants must have the information about the study, about what it expects of them, the capacity to understand that information, and the power to freely refuse consent. If a participant does not explicitly consent, they are not included in a study and, mostly, will not receive the intervention.

In the context of research with young people and other vulnerable groups, we often assume the participants themselves are not competent to provide informed consent for their own participation in the study due to their age, intellectual capacity, or circumstances. As such, it is common to ask that researchers collect opt-in consent from a caregiver, such as a parent, or a professional exercising parental responsibility. This is often required in addition to the participant's own assent, except where it would not make sense to seek this (eg they are too young to understand what participation involves).

Opt-out assent

We define the alternative here as 'opt-out assent.' In this scenario, parents, guardians or professionals are informed that the research is taking place, and given the same information they would be for opt-in consent, but are informed that unless they explicitly state that they do not want the CYP to be a part of the study, the CYP will be able to participate. This is often, but not always, one component of the process, with consent/assent also sought from the participants themselves and from gatekeepers, such as the school.

Prevalence

It is difficult to estimate the prevalence of the two approaches to consent/assent described above. However, a review of trials funded by two what works centres working with CYP, the Education Endowment Foundation and What Works for Children's Social Care, suggests that both are in wide current use, sometimes within the same study.

Over the next several sections of this chapter, we will consider in turn the ethical considerations relating to consent and assent; considerations that relate to the robustness of the research itself; and the practical considerations.

Ethical Considerations

We would like to emphasise at the outset that opt-in, informed consent should be the default setting in research involving human subjects, and particularly vulnerable groups. It should be the responsibility of funders, researchers, and those delivering interventions to justify not seeking informed, opt-in consent.

However, as we outline below, this default does not imply that there are no situations under which the alternative can be justified. In fact, it is possible that in many studies opt-out assent may be preferred. Nonetheless, the case needs to be made – and subject to independent ethical scrutiny and monitoring. As with all research design decisions, it is important that we do not mark our own ethical homework. Throughout the rest of this section, we consider the arguments that might be presented to support opt-out assent in a study with CYP.

Ethical counterfactual

An important consideration in social policy research is the idea of an ethical counterfactual. Just as trials attempt to use randomisation to understand what would have happened to treated individuals had they not been treated, we need to actively consider what would happen were this study not to take place.

In medicine, from which we develop most of our understanding of randomised trials, the answer is clear – the intervention would simply not be used in practice (Kuther, 2003). In social policy, this is usually not the case. Let us consider, for example, the case of 'Achievement For All', a whole-school intervention that aimed to improve students' grades. The intervention was successful, in that it was widely taken up, and deployed in hundreds of schools, before an RCT was conducted (Humphrey et al. 2020). This large-scale, robust trial found reliable and consistent evidence that, far from being beneficial, 'Achievement For All' was actively harmful to students' attainment.

As this example illustrates, in social policy, the counterfactual to a trial taking place is often not that nothing happens. Instead, it is that the intervention continues to be dumped, willy-nilly, into the water supply of the policy domain, potentially wasting money, or even harming thousands of people. In this context, it is legitimate to ask not just whether it is ethical to make use of opt-out assent, thus allowing the study to take place robustly, but also whether it is unethical to prevent that study from taking place.

A model of research ethics that assumes that any intervention to be studied is novel to the study and must be treated as such runs the risk of placing too high of an ethical burden on researchers and unintentionally preventing studies from being conducted at all. The burden on those who develop and deliver interventions in social policy should also be considered: unlike in medicine, where a drug cannot be authorised by regulators without a high-quality evaluation, developers in policy settings are under no obligation to have their intervention tested at all and many, confronted by high barriers to research implementation, will simply choose not to.

Consent counterfactual

As well as the ethical counterfactual that we've just discussed, it is also helpful to consider what people's expectation is of what they will be asked to provide consent for, compared to what decisions professionals are expected to make. Again, the contrast to medicine is a useful one. We would not expect a child to be given a drug, or any medical procedure, without the parents/guardians knowing about it or approving it, except under a rule of rescue: to do otherwise would be well outside of the norms of healthcare in the developed world. In other domains, the expectation of consent is much lower.

In schools, for example, students are told about higher education; about further education and vocational options; and given career advice. Parents' consent neither to this advice occurring, nor to the way in which it is given, or what their children are specifically advised to do. A huge range of activities – for example, what the Education Endowment Foundation describes as 'Teacher Moves' are chosen by individual teachers in their classrooms day after day, without parents being informed at all. For example, a teacher can choose to teach maths using Lego rather than in the abstract – or vice versa. A school can also change the menu in the cafeteria without consent. A teacher or a school is free to change the way in which they do these things without informing parents. In the case of a trial, in which a change occurs, and parents are made aware of its possibility in advance, and asked to voice concerns if they have any, far more consent is being sought than would ordinarily occur.

Intrusiveness of the research

Related to the above, we should look at how much the research activities intrude on participants' lives when considering the appropriate consent/assent regime. Many interventions involve a significant change to the participant's daily pattern; for example, withdrawing them from classes for additional support, or requiring them to take part in additional activities outside of school.

Summer schools to improve access to higher education, or schemes involving mentorship by sports coaches, similarly represent a change to what CYP would be doing otherwise. Where interventions are this intrusive and disruptive to a person's day, there is a clear need for opt-in consent. However, in all these cases, opt-in consent is unavoidable because participating young people will not simply be attending school in the usual place at the usual time.

Other interventions – like providing financial incentives for students (Sibieta et al., 2014), or extra tuition during school time and in school buildings (Torgerson et al., 2016) – are intrusive enough to warrant opt-in consent but do not mechanically require it (that is, the activity could physically take place without parental consent). Other activities – like receiving a letter from a

former student, or attending an assembly (Sanders et al., 2018) – are less intrusive, and are in many ways indistinguishable from variation in 'business as usual' for young people.

In many cases, even fairly substantial variations in practice and participants' daily patterns are either implicitly or explicitly delegated to other professionals such as teachers or social workers, who have responsibility for the education or wellbeing of the child. If the intervention itself is unlikely to be experienced as unusual or intrusive by participants or doesn't require them to spend time differently to how they usually would, the case for opt-out assent is more justifiable.

Impost on partners

When conducting research with institutional partners – be they schools, universities, or local authorities – we must also consider the burden of our research on those partners. Collecting consents from large numbers of people, and flexing delivery of an intervention around the potentially large number who do not consent for reasons having to do with inattention rather than a conscious desire, is administratively burdensome and requires delivery organisations to redirect their time and effort away from other activities, such as supporting students, in order to facilitate the research. In this case, the ethics of consent must also be weighed against the cost of the administrative burden of consent, particularly when this has an ethical element as regards what that administrative effort would otherwise have been used for.

Fairness of excluding people who want to join

If a potential participant has agreed that they would like to take part in the research, but their parent/guardian hasn't consented, then that participant would be excluded from some or all of the activities associated with the research. From the participant's perspective, this might mean they have to sit in a separate room and do a worksheet or other task while other pupils get to do a novel, additional activity. As we discuss later, CYP who experience this are more likely to be educationally disadvantaged, and the experience could contribute to a feeling of educational exclusion, as they may have had multiple experiences of missing out on activities because their parent/guardian failed to return the permission slip.

This could be mitigated by allowing young people whose parents don't consent to still take part in some aspects of the research; for example, joining activities but not being surveyed afterwards. However, whether this is possible depends on resources available to provide the intervention activities, and the nature of the activities themselves – including whether the 'activity' components can be separated from 'research' components and whether there are components it is appropriate for CYP to take part in absent parent consent.

Disclosure to parents/guardians of child's activities

Finally, we should consider whether there are limits to parents' right to know everything that happens with their children. Some interventions might relate to specific groups of young people, and the eligibility criteria, or the nature of the intervention, may risk disclosing to parents something about their child that they may not want known. This is the case for interventions that might relate to understanding a young person's sexual health or activity; their gender identity or sexual preference; their mental health; or other factors. We must

consider whether young people have the right to do some things that their parents do not approve of or know about.

Considerations for robustness of research

In this section, we move away from ethical arguments per se, to consider the effect of consent/assent on the robustness of studies. Although these considerations are more technical than those considered previously, they remain relevant when we consider the ethics of a study design: both because of the importance of facilitating research that has the capacity to improve public service provision for the better, and because it is unethical to ask for participants' time and information in service of research that is not robust.

Recruitment and statistical power

Multiple, large-scale studies have demonstrated that sample sizes can be significantly affected by using opt-in consent procedures in school-based research with CYP. Experimental trials have shown that participation rates can drop from 79 per cent to 29 per cent, (Courser et al., 2009) or 96 per cent to 41 per cent (Spence et al., 2015), when opt-in consent is used instead of opt-out to recruit students in survey-based research. Another survey-based study in urban, ethnically diverse middle schools in the US reported a 31 per cent participation rate when using opt-in consent (Anderman et al., 1995). A 2010 analysis of different consent processes used in school-based research found that opt-out assent procedures resulted 'in parental permission from 90-100 per cent of eligible students' while opt-in consent resulted in 'parental permission from 30-60 per cent of students.' (Secor-Turner et al., 2010). Similarly, significant differences can be found when comparing online survey respondents' opt-in and opt-out rates when asked if they would like to receive future communication (Bellman et al., 2001).

It is therefore likely that half the potential sample will be lost if opt-in consent procedures are required vs. opt-out assent. For some studies this will not be an issue, if the sample size is still sufficient to detect the expected effect size, or there is sufficient budget to over-recruit to account for non-consent. However, for many studies, this level of attrition would be challenging, especially given the time and financial cost of engaging more schools in order to compensate.

If opt-out assent is not possible, researchers would then be faced with the choice of either substantially increasing the budget (which may not be possible) or accepting that their research will likely be underpowered. It is our view that conducting research that is known before the fact to be statistically underpowered is ethically problematic, as any burden on participants occurs without a high chance of the research question being answered. A similar argument has been made elsewhere; for example, Menzies et al. (2016). In this study, the majority of schools took an opt-out approach, while one school took an opt-in approach. This resulted in lower sample sizes, and the ethics committee at the University of Durham concluded that the small sample size meant that this school could not ethically be included in the study.

Representativeness

Even if the budget permits over-recruitment to compensate for non-consent, there is still reason to be concerned about the robustness of school-based research with opt-in consent.

When opt-in consent is used in school-based research, there is evidence that those parents/guardians who engage are systematically different from those who do not. A study that administered surveys to students using opt-in consent one year, and then opt-out the next exemplifies this: the least deprived parents were 17 times more likely to assent via an opt-out process, compared to opt-in consent, while parents from the most deprived groups were 13 times more likely (Spence et al., 2015). Students with antisocial and substance use behaviours have also been found to be underrepresented when opt-in consent is used (Courser et al., 2009).

The differences between samples recruited using opt-in and opt-out procedures is also evident in studies where researchers first attempt opt-in consent but then default to assuming assent when there is no response. Students who display problem behaviours, are older, have lower academic performance, do not live with both parents, are male, are non-white, have less educated parents, have higher levels of absenteeism, or have special education provision are found to be underrepresented in opt-in consent samples to a statistically significant degree (eg Shaw et al., 2014; Went et al., 1993; Henry et al., 2002; Anderman et al., 1995; Esbensen et al., 1999). The differing rates of opt-in consent across ethnic groups can be particularly pronounced, with non-white students opting in at a rate approximately 13 percentage points below their white peers (Anderman et al., 1995; Esbensen et al., 1999). Consistently, those that do less well in educational settings than their socially advantaged peers are underrepresented when opt-in consent procedures are used.

Therefore, school-based studies using opt-in consent procedures are limited in their ability to make inferences about the impact of the interventions on marginalised and disadvantaged students. This is extremely problematic, given that these students are arguably those we should be most interested in finding effective support interventions for, as well as identifying any interventions that may be unexpectedly harmful.

Practical considerations

Data protection

Data protection in the UK is governed by the UK General Data Protection Regulation (UK GDPR). Under UK GDPR, there are four legal bases for processing personal data, which includes any data that could be used to identify a research participant. One of the legal bases is consent (Hoonagle et al., 2019). However, it is important to distinguish between GDPR consent and ethics consent.

Very few researchers rely on GDPR consent as the legal basis for processing personal data. It is far more common for researchers to rely on 'task in the public interest' as the legal basis. Under this legal basis, opt-in consent for data processing is not required, as long as the researcher can demonstrate that they have made all reasonable efforts to make data subjects aware of the ways in which their data is being processed and advised them how they may withdraw their data.

Research ethics consent is, therefore, separate from GDPR consent, and it is possible to access personal data, including administrative data such as the National Pupil Database, without obtaining opt-in consent.

Budget implications

With opt-in consent, most projects would likely need to commit additional budget to the consent process. A variety of incentive systems involving both students and staff, multi-pronged communication to reach parents, talks at school events, and even school visits by researchers to directly encourage students to return consent forms have been used to improve engagement when using opt-in consent procedures. (Henry et al., 2002; Ji et al., 2004; Esbensen et al., 2008; Secor-Turner et al., 2010). These methods are successful insofar as they achieve participation rates that are comparable to opt-out procedures (between 70 per cent to 95 per cent).

However, the cost involved in maintaining these processes is significant. Existing research suggests that costs can range from between approximately US$1000 per school for less intensive methods – as in Esbensen et al. (2008), to $20-$25 per student for more intensive methods as in Tigges (2003), in a non-school context. Secor Turner et al. (2010) achieved very high active consent rates in urban middle schools and estimated their costs to be $11, and 25 minutes of research staff time, per consent form returned. If this were applied to the context of a study in the UK involving around 50 schools, these figures would imply an additional cost of over £400,000, For non-school domains, where participants may be more vulnerable, or less easily reached, this cost could be higher still.

It is also worth noting that even when these techniques are employed, some systematic bias often remains in the resulting sample. For example, students with poor grades (Unger et al., 2004) or higher rates of absenteeism (Secor-Turner et al., 2010) can be less likely to engage in opt-in procedures, as can schools with larger classes (Esbensen et al., 2008), even when additional engagement techniques are used.

Feasibility for schools and other partners

Requiring opt-in consent may be unmanageably complex for delivery partners such as schools. For instance, they may be required to send and resend forms, maintain lists of consents returned, follow up non-responders, arrange events for parents to learn about the research, and arrange for researchers to attend the schools to speak to parents. All of these activities require additional work from school staff that can be burdensome and impose on the relationship between the school and its community. Assuming that opt-in consent is obtained from parents/guardians for around half of a class, the school is then required to keep track non-consenters, and arrange alternative activities and possibly space for them while the research is occurring.

This level of coordination complexity can put schools off participating in research, making sample size considerations more challenging, and reducing the external validity of the study. In some instances, researchers may be able to assume some of the burden, working with school administrators to coordinate alternative activities and space, but this is not always feasible or appropriate; for instance, researchers will usually not be able to contact parents directly or organise events on school grounds without school staff present.

Compensating for assent

As we emphasised earlier in this chapter, the default for all research studies should be fully-informed opt-in consent, and that opt-out assent must be justified. As a part of this, it is important to recognise that a safeguard on CYP is removed if we remove the requirement for parents/guardians need to provide opt-in consent. As researchers, we are therefore obligated to ensure that our processes are robust enough to ensure that the CYP remain safe. Whether these processes are, or can be, robust enough to allow for us to ethically seek opt-out assent rather than opt-in consent, depends not just on the nature and implementation of the processes themselves, but also on the level of risk associated with the study. A routine classroom intervention to improve grades in maths is clearly less risky than one that relates to a young person's mental health.

For example, evaluators should provide participants and their parents/guardians with multiple opportunities to opt out at each data collection point and work with schools and other partners to ensure that both parents and pupils are fully informed. An information sheet should be disseminated to parents and guardians before each round of activity.

Where possible, we should utilise multiple forms of communication, such as asking schools to include briefings about the research at parents' evenings. This could be followed up with a post-experimental debrief sent to all participants' parents, reminding them of the research and its purpose, and giving them a further opportunity to opt their child out. It should be made clear to both parents/guardians and CYP at every stage that they are free to withdraw without giving a reason and that doing so will not impact their standing. Delivery partners should be regularly briefed by researchers about the importance of parents/guardians and CYP not perceiving pressure to assent to the research if they do not wish to.

It is important that evaluators, researchers, researchers, and commissioners (including what works centres and governments) work with partners to develop appropriate communication channels that maximise parental engagement in their contexts. Where a study relies mainly or mostly on CYP's willingness to participate, researchers should work with professionals to develop age-appropriate materials to give CYP information about the research. These professionals themselves should also have it made clear to them that they have a heightened responsibility in the absence of parental opt-in consent.

Researchers, evaluators, and delivery partners must have safeguarding procedures that are fit for purpose, with clear monitoring processes to identify risks to participants and transparent reporting lines outlining responsibilities in case specific issues arise.

Finally, it is important to consider the composition of the research team itself. If we are concerned (as we are), that research that relies on opt-in consent excludes marginalised and disadvantaged groups, we must take inclusion seriously in all aspects of our research. This means actively endeavouring to ensure that our research teams are diverse and representative of the population involved in our studies; and ensuring that we make greater use of collaborative research approaches that involve and engage those with lived experience in the design of both interventions to be evaluated, and the evaluations themselves.

Discussion

In this chapter we have discussed the use of opt-in consent and opt-out assent in randomised trials in social policy relating to children and young people, although we argue that the implications are wider to a range of social policy contexts.

As we have raised, opt-in consent should be treated as the default for all studies, with justification and mitigation needing to be presented to an independent ethics committee prior to a decision being taken about whether opt-out assent is permissible in the specific context of the study.

This being said, we believe that there are a number of contexts in which opt-out assent for trials involving CYP is valid, particularly where interventions are minimally burdensome or intrusive; where CYP themselves have a say in whether they participate; where interventions take place at a whole-class or whole-school level, and where there are professionals with some responsibility for the care of CYP involved in the trial.

Where these conditions are met, opt-out assent seems to be justified, so long as steps are taken to ensure the safeguarding of participants by researchers and other professionals who are involved in their lives.

Conversely, there can be, to our mind, important ethical arguments against opt-in consent for certain studies, if it harms the robustness of a trial through lower participant numbers (thus invalidating the ethical argument for the research itself), and, more importantly, through reducing the representativeness of the research in particular by reducing the participation of disadvantaged and underrepresented groups.

We hope that this chapter provides a useful summary of the issues and arguments for and against opt-out assent in trials and can be used to reduce between-study heterogeneity (and hence increase generalisability) in how trials are conducted when working with young or otherwise vulnerable people.

References

Anderman et al. (1995) Selection bias related to parental consent in school-based survey research. Evaluation Review, 19(6)

Bonetti, S. (1998). Experimental economics and deception. Journal of Economic Psychology, 19(3), 377-395.

Courser et al. (2009) The impact of Active Consent Procedures on Nonresponse and Nonresponse Error in Youth Survey Data: Evidence from a New Experiment. Evaluation Review, 33(4)

Edovald, T., & Nevill, C. (2021). Working out what works: The case of the Education Endowment Foundation in England. ECNU Review of Education, 4(1), 46-64.

Esbensen et al. (1999) Differential attrition rates and active parental consent. Evaluation Review, 23(3)

Esbensen et al. (2008) Active parental consent in school-based research: how much is enough and how do we get it? Evaluation Review, 32(4)

Kelman, H. C. (2017). Human use of human subjects: The problem of deception in social psychological experiments. In Research Design (pp. 189-204). Routledge.

Kuther, T. L. (2003). Medical decision-making and minors: issues of consent and assent. Adolescence, 38(150), 343.

Henry et al. (2002) The effect of active parental consent on the ability to generalize the results of an alcohol, tobacco, and other drug prevention trial to rural adolescents. Evaluation Review, 26(6)

Hoofnagle, C. J., van der Sloot, B., & Borgesius, F. Z. (2019). The European Union general data protection regulation: what it is and what it means. Information & Communications Technology Law, 28(1), 65-98.

Humphrey, N., Squires, G., Choudry, S., Byrne, E., Demkowicz, O., Troncoso, P. & Wo, L., (2020) Achievement for All Evaluation.
https://d2tic4wvo1iusb.cloudfront.net/documents/pages/projects/Achievement_for_All_final.pdf?v=1673013925.

Menzies, V. and Hewitt, C. and Kokotsaki, D. and Collyer, C. and Wiggins, A. (2016) 'Project Based Learning: evaluation report and executive summary.', Project Report. Education Endowment Foundation, London.

Miller, S., Davison, J., Yohanis, J., Sloan, S., Gildea, A., & Thurston, A. (2017). Texting Parents: Evaluation Report and Executive Summary. Education Endowment Foundation.

Sanders, M., Burgess, S., Chande, R., Dilnot, C., Kozman, E., & Macmillan, L. (2018). Role models, mentoring and university applications-evidence from a crossover randomised controlled trial in the United Kingdom. Widening Participation and Lifelong Learning, 20(4), 57-80.

Sanders, M & Breckon, J (2023). The What Works Centres. Policy Press
Sanders, M., Mitchell, C., & Ni Chonaire, A. (2020). Effect sizes in education trials in England. Available at SSRN 3532325.

Secor-Turner et al. (2010) Active parent consent for health surveys with urban middle school students: processes and outcomes. Journal of School Health, 80(2)

Shaw et al. (2014) Bias in student survey findings from active parental consent procedures. British Education Research Journal, 41(2)

Sims, S., Anders, J., Inglis, M., & Lortie-Forgues, H. (2020). Quantifying 'promising trials bias' in randomized controlled trials in education.

Spence et al. (2015) Does the use of passive or active consent affect consent or completion rates, or dietary data quality? Repeat cross-sectional survey among school children aged 11-12 years. BMJ Open

Stokes, L. (2022). Supervision for Designated Safeguarding Leads in schools: focus on child sexual abuse- Information Sheet for schools.

Tigges, B. (2003) Parental consent and adolescent risk behavior research. Journal of Nursing Scholarship, 35(3)

Torgerson, C., Ainsworth, H., Buckley, H., Hampden-Thompson, G., Hewitt, C., Humphry, D., ... & Torgerson, D. (2016). Affordable online maths tuition: evaluation report and executive summary. Education Endowment Foundation.

Unger et al. (2004) Characteristics of adolescents who provide neither parental consent nor refusal for participation in school-based survey research. Evaluation Review, 28(1)

Went et al. 1993) Demographic, psychosocial and behavioral differences in samples of actively and passively consented adolescents. Addictive Behaviors, 18(1)

Chapter 10. Multiple Comparisons in RCTs

Introduction

Often, when we are conducting a randomised controlled trial, we are interested in testing more than one hypothesis. For example, we might want to understand the effect of a schools-based intervention on both educational attainment, and progression to higher education. This makes a lot of sense, because many adverse outcomes share common causes, and so we might expect effective solutions to one, to also improve the other. Moreover, we might not merely be interested in the average effect for the whole sample but instead wish to consider multiple subgroups. An example of this might be that the Education Endowment Foundation always measures effects for both the whole sample, and the portion of that sample which is eligible for free school meals. As discussed in Chapter 8 there is a strong argument for considering more carefully the equity and equality impacts of randomised trials, alongside a more general – and understandable – wish to determine whether they are different effects for subgroups.

When we measure multiple outcomes, or effects for multiple groups, and conduct statistical tests for each, we are making multiple comparisons. As we've shown, this is a legitimate thing to want to do, but it is not without issues. Specifically, the more comparisons we conduct, the higher the probability of making a Type I error, or false positive. The more tests that are performed, the higher the likelihood that a seemingly significant effect will be simply the result of chance. In order to guard against this effect, there are necessary statistical adjustments that need to be made depending on the number of comparisons and the design of the trial itself. This report discusses the nature of this issue, explores methods for correcting it, and provides a detailed explanation of the methods available for controlling the false discovery rate (FDR) while preserving statistical power.

The Problem of Multiple Comparisons

When conducting statistical significance tests, the typical threshold probability of a false positive result is 5 per cent, also known as the probability of a Type I error. This means that had we tested the hypothesis with 100 new sets of data that only differed due to random error, 5 out of 100 tests would be expected to yield a significant result purely by chance. However, as the number of tests increases, the likelihood of obtaining at least one false positive grows. This phenomenon can be analogised to rolling dice: while the probability of rolling a six on one throw is 1/6, the probability of rolling at least one six in two consecutive rolls increases to 11/36, which is much higher than the original 1/6. Similarly, when performing independent tests with a 5 per cent false positive rate, the probability that at least one test is a false positive increases beyond 5 per cent.

For instance, when running two independent tests, the probability of both being false positives is only 1/400. However, the probability that at least one of them is a false positive rises to about 9.75 per cent. Thus, conducting numerous tests increases the likelihood of obtaining a spurious result, which necessitates the application of correction methods to adjust for multiple comparisons. This is important, because otherwise we might end up falsely finding that an intervention is effective, and subsequently waste public money on that intervention.

Correcting for Multiple Comparisons

Several approaches have been developed to address the issue of multiple comparisons. The goal of these corrections is to control the overall error rate across the tests being conducted, ensuring that the probability of a false positive remains at the desired level (eg, 5 per cent).

1. **Bonferroni Correction:** The Bonferroni correction (Bonferroni, 1936) is one of the simplest methods used to adjust for multiple comparisons. In this approach, the threshold for statistical significance (alpha) is divided by the number of tests (n). For example, if two tests are conducted with an α level of 0.05, the corrected threshold becomes 0.05/2 = 0.025 for each test. Similarly, if five tests are performed, the corrected level is 0.05/5 = 0.01.

While this method is easy to implement, it is highly conservative and may reduce statistical power, especially when a large number of tests are performed. This can lead to an increased likelihood of Type II errors (false negatives). In general, this procedure is recommended when:

- A 'universal null hypothesis' test that all tests are not significant is required,
- It is imperative to avoid a type I error – which is rarely the case in public policy or public administration trials.
- A large number of tests are carried out without preplanned hypotheses – which can be avoided via pre-registration.

2. **Benjamini – Hochberg's Step-Up Procedure:** Hochberg's step-up procedure (Benjamini and Hochberg, 1995), offers a more 'powerful' alternative to the Bonferroni correction – in that it controls the false discovery rate while preserving statistical power. Contrary to the Bonferroni correction which assigns the same threshold to all tests, in order to perform Benjamini – Hochberg's Procedure, we rank p-values from smallest to largest and compare them with a sequence of values accepting the alternative hypothesis for all smaller or equal p-values to that of any one found less than it's critical value. Utilising the formula Pkalpha*kn where n is the total number of tests and k the highest ranked p-value for which we accept the alternative hypotheses.

Steps: One starts by examining the largest p-value P(n), if P(n) ≤alpha, then all alternative hypotheses are accepted; if not, then move to P(n-1), if P(n-1)≤ alpha*(n-1)/n, then all n-1…1 hypotheses are considered statistically significant, if not, continue until statistical significance is found for some ranked test, after which all smaller ranked p-values are identified as statistically significant.For example, consider a scenario with five hypothesis tests and a 5 per cent significance level.

- H1: 0.04
- H2: 0.06
- H3: 0.2
- H4: 0.015
- H5: 0.005

We can see how both methods compare against each other in the following table:

Table 1 Benjamini-Hochberg vs Bonferroni, Multiple comparisons correction

			Benjamini – Hochberg		Bonferroni	
Rank (k)	**Hypothesis test**	**P-value**	**Threshold (alpha*k/n)**	Ha:μ≠0	**Threshold (alpha/n)**	Ha:μ≠0
1	P(n-4):H5	0.005	≤0.01	**Accept**	≤0.01	**Accept**
2	P(n-3):H4	0.015	≤0.02	**Accept**	≤0.01	Reject
3	P(n-2):H1	0.04	≤0.03	Reject	≤0.01	Reject
4	P(n-1):H2	0.06	≤0.04	Reject	≤0.01	Reject
5	P(n):H3	0.2	≤0.05	Reject	≤0.01	Reject

Looking at the table we see that P(n-3) is the largest p-value for which we reject the null and accept the alternative, this implies that hypotheses' 1 and 2 ranked p-values are then considered statistically significant. It is important to note that if instead P(n-4)>0.01 and still P(2)≤0.02, both hypotheses would still be accepted.

As it is evident from the table above the Bonferroni correction applies a high premium to multiple comparisons resulting in more rejections. Instead, we recommend the use of Benjamini – Hochberg's step-up procedure, which preserves more power while still ensuring an acceptable false-discovery rate, making it more appropriate when researchers want to balance the need to avoid false positives with the desire to maintain statistical power. This is consistent with the statistical analysis guidance of several WWCs (CHI, 2024; TASO, 2023; WWCSC, 2021).

Applying corrections for Multiple Comparisons

The decision to apply multiple comparison corrections should be based on the number of tests being conducted and the nature of the outcomes being examined. It is advised to correct for multiple comparisons within a category of outcome, such as primary or secondary outcomes, particularly when a large number of comparisons are made.

The number of comparisons is determined by the formula:

$$\text{Comparisons} = (\text{Arms} - 1) \times \text{Number of outcomes}$$

For instance, in studies with multiple treatment arms and outcomes, the number of comparisons grows quickly, and corrections become crucial to ensure valid results. The use of a Benjamini – Hochberg step-up procedure is recommended when dealing with an especially high number of comparisons within a category of outcome.

The box below indicates in what instances we recommend using multiple comparison adjustments. The shaded cells in the table below show when a Hochberg step-up procedure should be used in line with this guidance.

Table 2 Multiple comparison adjustment by numbers of trial arms and outcomes

	Number of outcomes within a category			
Number of trial arms	**1**	**2**	3	**4**
2				
3				
4				
5				
6				

Implications for Sample Size Calculations

The typical sample size calculation formula for the common t-test either in the case of the individual level trial or in the case of the design-adjusted cluster randomised trial, sets critical z-values (or t-values depending on sample size) corresponding to the desired significance (alpha) and power (beta) detection levels of the test. Because both methods for correcting for multiple comparisons reduce the individual test's alpha, achieving statistical significance becomes more difficult when comparing multiple outcomes, which in turn requires a larger sample size to maintain adequate statistical power. See Annex 2 for a breakdown of sample size calculations under different scenarios and adjustments for multiple comparison methods.

Bonferroni Correction: In the case of Bonferroni, for k number of comparisons and a pre-specified significance level alpha, the new alpha* required in the sample size calculation becomes alpha*=alpha/n. So, if the trial has two arms and has one outcome measure, that's one comparison, and so the threshold for significance is $p<0.05$ as normal. If we have two arms and two outcome measures, that gives us two comparisons, and so the threshold for significance becomes 0.05 divided by 2, which is equal to $p<0.025$. If we are making five comparisons, then the threshold becomes 0.05 divided by 5, which is $p<0.01$.

As a simple rule of thumb, to achieve a significance level of 5 per cent across n outcomes, we would need the same sample size to increase by approximately 10% for every additional comparison we're making. Taking the previous example where we make five comparisons, this implies that power calculations would need to be run for a new alpha of 0.05/5=0.01, or an increase of almost 50 per cent in the sample size requirement.

Benjamini – Hochberg's Step Up procedure: The impact on power when using Benjamini – Hochberg's Step Up procedure is less severe compared to Bonferroni because the adjustment is less conservative, since the method focuses on controlling the family-wise error rate (FWER) but does so with more flexibility. Since adjustment of p-values in this procedure is determined by ranking them, there is no closed-form formula that can solve for sample size. Researchers can instead use the Bonferroni Correction to inform their power calculations and then perform the Benjamini – Hochberg Procedure when considering significance. Alternatively, available software packages that calculate power for Benjamini – Hochberg's and other FDR adjustment procedures can also be used (see 'pwrFDR' in R).

Conclusion

Multiple comparisons increase the risk of false positives in statistical analyses, necessitating the use of correction methods to control for Type I errors. While the Bonferroni correction provides a simple approach, it can be overly conservative, leading to a loss of statistical power. Benjamini – Hochberg's step-up procedure, on the other hand, offers a more balanced alternative, controlling the false discovery rate without excessively compromising the ability to detect true effects, allowing for more power. When researchers are considering applying Multiple Comparison methods, they should carefully consider the number of hypothesis tests they want to conduct as part of their analysis as well as the structure of the study itself to ensure valid and reliable results.

References

Bonferroni, C. E., 1936: Teoria statistica delle classi e calcolo delle probabilità. Pubblicazioni del Regio Istituto Superiore di Scienze Economiche e Commerciali di Firenze, Vol. 8, 3–62.

Benjamini, Y., and Hochberg, Y., (1995). 'Controlling the false discovery rate: a practical and powerful approach to multiple testing'. Journal of the Royal Statistical Society, Series B. 57 (1): 289–300

Part 3
Beyond the RCT

Sometimes you need to go beyond the standard models of RCT that we have laid out in part one. Part three of the book can be thought of as the most 'experimental', describing types of trials that are relatively less tested in social policy contexts, or providing data tables that can be useful in some trial designs. This part of the book is more specialised – with trial designs that are only likely to be useful under quite narrow circumstances – but nonetheless which a committed evaluator or commissioner might wish to familiarise themselves with. This part contains chapters on;

- Complex trials — how to approach trials when interventions or contexts limit the usefulness of more straightforward designs.
- Megastudies – trials that use large samples to test large numbers of different treatments at once – in general most useful for communications trials or tests of light touch interventions
- Split plot trials — which test multiple interventions simultaneously while making use of randomisation at multiple levels of clustering.
- Winner stays on trials – in which multiple interventions are tested sequentially rather than simultaneously, in an adaptive trial design.
- Useful parameters for trials in higher education and homelessness, which can be used to design more efficient randomised trials in those policy domains.

Chapter 11. Intracluster correlation rates in policing, higher education and homelessness

Introduction

There has been a notable rise in the use of randomised controlled trials to assess different interventions in crime, higher education (HE) and homelessness, facilitated in part through the establishment of the Centre for Transforming Access and Student Outcomes in Higher Education (TASO) and the Centre for Homelessness Impact (CHI). In this chapter, we apply the general principles and processes of good trial design -in particular paying attention to power calculations and intracluster correlation (ICC) rates-to the specific use cases of running trials in policing, HE and homelessness.

The purpose of this chapter is to provide a series of parameters which could be usefully used by evaluation designers or commissioners when designing trials in these three areas. Providing these sets of intra-cluster correlations is therefore useful in designing cluster randomised trials to ensure that they are well powered.

Context of policing

One of the major methodological innovations of the 21st century in policing and crime research has been the rise of the use of randomised controlled trials to identify the impacts of particular policies and practice (Torgerson, 2008). Although the canonical randomised controlled trial would see randomisation occurring at the level of an individual observation — for example, an individual crime or an individual interaction between police and civilians, for many interventions, this idealised trial is not possible.

Common evaluations in policing might focus on training for officers LaMontagne et al (2016), innovations in policing (Quinton, 2011, Macqueen & Braford, 2015), or 'hot spot' policing (Williams and Coupe, 2017). In these cases, randomisation takes place at the level of a cluster — that is, a higher level which contains individual operations within it. A common level of randomisation is the individual officer or selection of officers — for example trials of body worn video cameras randomise officers either to have, or not to have, a camera Ariel et al (2016). Randomisation can also occur at a geographical level, with all of the officers operating within a geography, for example based at a particular station (Scantlebury et al, 2017); or all of the interactions within that area, following a given approach.

Context of higher education

Many interventions in HE widening participation do not lend themselves to individual randomisation. As an example, we conducted a trial investigating the effect of letters sent to high-achieving students from role models on increasing applications to selective institutions, (Sanders et al., 2017). For this,, saw randomisation was done at the school level, on the basis that the data could only be accessed at that level. Other interventions, such as mentoring and inspirational talks, where current university students visited local schools and colleges and gave inspirational talks while providing information on the costs and benefits of university (Sanders et al., 2018), could only plausibly be delivered at the level of the school. For many, perhaps most outreach activities, the school is the natural interlocutor and level of delivery for an HE institution while for many others, concerns about the ethics of excluding some people within the same school from an intervention or methodological challenges around spillovers and contamination, make the school a sensible level of randomisation.

Context of homelessness

Similarly to HE, testable interventions in homelessness do not typically lend themselves to individual-level randomisation. Particular districts can exhibit higher incidence as a result of area level factors: areas with a higher level of unemployment, crime rate, and poverty will often be correlated with higher numbers of households experiencing homelessness (Fargo et al., 2013; Mabhala et al., 2020). Recent UK based studies that aim to reduce homelessness, including Sanders and Picker (2023a), Sanders and Picker (2023b), and EDIT (2022), all rely, to some extent, on assignment of interventions at the level of a cluster, typically a geographical unit.

Statistical power and sample size calculations

In all randomised trials, power, or sample size calculations, area crucial component of the study's design. Unlike observational studies, in which researchers often must take data as a given, researchers conducting trials often have some degree of agency. Their task in these calculations is to determine how large the study needs to be – how many participants there must be – in order to have a good chance of detecting an effect size that is of interest, given the cost and resource intensity of the intervention. Even where a study's size is fixed by budget or other constraints, sample size calculations can tell us whether the study is likely to contribute much to the state of our knowledge and hence whether it is ethical to conduct an RCT at all. At the other end of the spectrum, HE providers, what works centres, and local authorities have finite financial and human resources, and spending more of them than needed on a research project represents a poor use of public money. Accurate power calculations, based on accurate assumptions, are therefore critical to ensuring that research is ethical, that it is able to answer the research questions that it sets out to and that it makes efficient use of scarce resources. In this chapter, we provide calculations of the intracluster correlations for key potential outcome measures for cluster randomised trials in Policing, HE participation and homelessness.

Datasets – Policing, HE and homelessness

Policing

The data used stem from data.police.uk, a site offering street-level and outcome data on crime and policing in England, Wales and Northern Ireland. They contain each crime identified and recorded within a 2011 Lower Layer Super Output Area (LSOA) for each constabulary in the UK and time point. In order to conduct the analysis, constabulary datasets of all street and outcome crimes were extracted and appended for each month, covering a 12-month period from January to December of 2022. Crimes that were not assigned to an LSOA were listwise dropped for each time point. Two auxiliary datasets were also extracted from the Office for National Statistics', the first identifying, among other geographic variables, frozen 2011 Census Lower Layer Super Output Areas (LSOA), Middle Layer Super Output Areas (MSOA) and current local authority districts (LAD) as of May 2022 in the UK. This allowed for the identification of the respective Middle Layer Super Output Area (MSOA) in which LSOAs belong. The second dataset identified resident population for Lower layer Super Output Areas (LSOAs) in England and Wales from 2020.

Counts were then separately computed at crime/outcome type and LSOA levels. Following crime statistics literature, crime rates were then calculated as the number of crimes per 1000 people. Street crimes were categorised to the following types:

- Anti-social behaviour
- Bicycle theft
- Burglary
- Criminal damage and arson
- Drugs
- Other crime
- Other theft
- Possession of weapons
- Public order
- Robbery
- Shoplifting
- Theft from the person
- Vehicle crime
- Violence and sexual offences

Similarly, outcome types were assigned to the following categories:

- Action to be taken by another organisation
- Formal action is not in the public interest
- Further action is not in the public interest
- Further investigation is not in the public interest
- Investigation complete; no suspect identified
- Local resolution
- Offender given a caution
- Offender given a drugs possession warning
- Offender given penalty notice
- Suspect charged
- Suspect charged as part of another case
- Unable to prosecute suspect

The Intra-cluster correlation estimates were computed using variance estimates generated by empty multilevel models with the natural log of crime rates per 1000 residents as the outcome variable, to improve normality. In order to examine clustering, three different clustering structures were assumed:

1. Crime/Outcome types nested in LSOAs

2. LSOAs nested in constabularies,

3. LSOAs nested in MSOAs

Higher education

To calculate ICCs in HE, the primary dataset used was of key stage 5 HE outcomes for schools in the UK. The data provided percentages of students progressing to HE, including more specific destinations such as Russell Group universities and Oxbridge. This dataset was combined with auxiliary data on the number of students in each school that belonged to the key stage 5 cohort. The number of students that progressed was then inferred and data were expanded at pupil level, generating a student-level binary outcome of progressed/not-progressed. This transformation allowed the estimation of ICCs for students nested in schools.

Homelessness

The dataset used was of the detailed local authority level homelessness tables (DLUHC, 2023), which provide, among other statistics, the number of households threatened with homelessness and the number of households owed a homelessness duty. For the latter, numbers are also categorised by number of households owed a duty by the following support needs:

- Young person aged 16-17 years
- Aged 18-25 years requiring support to manage independently
- Young parent requiring support to manage independently
- Care leaver aged 18-20 years
- Care leaver aged 21+ years
- Physical ill health and disability
- History of mental health problems
- Learning disability
- At risk of / has experienced sexual abuse / exploitation
- At risk of / has experienced domestic abuse
- At risk of / has experienced abuse (non-domestic abuse)
- Drug dependency needs
- Alcohol dependency needs
- Offending history
- History of repeat homelessness
- History of rough sleeping
- Former asylum seeker
- Old age
- Served in HM Forces
- Access to education, employment, or training

ICC Tables – HE, homelessness, and policing

Higher education

Table 1 ICCs of pupil progression

	Sustained level 4 or 5 destination	Oxford or Cambridge	Russell group institution	Top third HE destination	All HE destinations	Sustained apprentice-ships	Sustained level 4 or higher destination
All Schools	0.245	0.690	0.342	0.341	0.212	0.272	0.198
All Non-Selective Schools	0.171	0.492	0.243	0.232	0.157	0.292	0.158
All Selective schools	0.284	0.254	0.226	0.246	0.204	0.173	0.210
Sponsored academy	0.228	0.910	0.296	0.257	0.184	0.461	0.186
Academy converter – mainstream	0.171	0.457	0.275	0.273	0.183	0.232	0.188
Academy 16-19 converter	0.066	0.507	0.149	0.116	0.042	0.058	0.050
Agriculture and Horticulture College	0.075	NA	0.15	0.158	0.082	0.855	0.051
Art, Design and Performing Arts College	0.647	NA	0.215	NA	0.057	NA	0.079
City technology college	0.002	0.557	0.296	0.181	0.091	0.684	0.119
Community school	0.178	0.433	0.181	0.177	0.135	0.286	0.128
Free school – mainstream	0.661	0.937	0.527	0.561	0.764	0.947	0.771
Free school – 16-19	0.420	0.837	0.561	0.655	0.280	0.170	0.334
Foundation school	0.144	0.506	0.235	0.262	0.133	0.286	0.136
General FE College	0.088	0.980	0.233	0.202	0.092	0.263	0.066
Independent school	0.846	NA	0.938	0.950	0.966	NA	0.967
Sixth Form College	0.154	0.699	0.235	0.232	0.100	0.15	0.109
Studio school	0.546	NA	0.293	0.498	0.348	0.646	0.149
University Technical College	0.116	NA	0.237	0.234	0.168	0.223	0.085
Voluntary aided school	0.280	0.588	0.233	0.260	0.204	0.368	0.209
Voluntary controlled school	NA	0.338	0.056	0.097	0.049	0.053	0.049

The results indicate large heterogeneity between estimates for school types as well as between destinations. ICCs for all schools are indicative of a positive correlation between HE ranking and ICC, with Oxford or Cambridge generating the highest at a value of 0.690 and Top 3rd and Russell Group destinations the second highest.

This pattern is repeated when looking at individual school types, with Oxbridge estimates quite high, particularly for specific school types. The reason for this pattern observed could be related to the fact that a select number of schools may have a much higher relative propensity to send students to higher ranking universities. This, in conjunction with other schools only sending small numbers, would in turn translate to larger heterogeneity of school-specific progress, inflating ICCs. Moreover, it is important to note that clusters for higher-ranking universities were more imbalanced, which would suggest wider confidence intervals for the ICCs estimated. Particularly high ICCs are also observed for independent schools. Estimated ICCs for selective schools are less variable between destinations; nonselective schools, on the other hand, echoed the pattern observed earlier, with Oxbridge exhibiting the highest ICC value.

Estimates by region are also provided below. The average range between destinations is around 0.2 to 0.4, with the exception of Oxbridge for which values are again quite high.

Homelessness

Estimated ICCs are shown in the tables below, calculated for each region by year for households threatened with homelessness (Table 1) and households experiencing homelessness (Table 2).

Table 2a ICCs by region and year of households threatened with homelessness

	2022-2021	2021-2020	2020-2019	2019-2018
All	0.090	0.134	0.084	0.100
East Midlands	0.084	0.089	0.070	0.090
East of England	0.076	0.104	0.079	0.091
London	0.127	0.131	0.096	0.096
North East	0.077	0.128	0.134	0.104
North West	0.094	0.170	0.111	0.080
South East	0.062	0.134	0.058	0.112
South West	0.075	0.189	0.064	0.094
West Midlands	0.108	0.112	0.071	0.142
Yorkshire and The Humber	0.077	0.108	0.073	0.047

Table 2b ICCs by region and year of households experiencing homelessness

	2022-2021	2021-2020	2020-2019	2019-2018
All	0.120	0.105	0.102	0.096
East Midlands	0.097	0.119	0.118	0.096
East of England	0.070	0.077	0.082	0.069
London	0.274	0.115	0.077	0.076
North East	0.070	0.085	0.072	0.065
North West	0.091	0.118	0.119	0.120
South East	0.111	0.089	0.105	0.098
South West	0.116	0.123	0.103	0.098
West Midlands	0.149	0.102	0.091	0.095
Yorkshire and The Humber	0.093	0.101	0.097	0.102

Overall, estimates range around an ICC of 0.1 to 0.2, with the largest values observed for the region of London followed by the North and South West, suggesting stronger clustering effects compared to other regions. There is a faint indication of some increase in the estimates as we move toward the more recent dates, with 2020-21 year exhibiting the highest estimates over all regions for households threatened with or experiencing homelessness.

Tables 3 and 4 contain ICCs by support needs and region, estimated for the most recent years of 2022-2021 and 2021-2020.

Table 2c Household homelessness ICCs by support needs and region (2021-22)

	All	East Midlands	East of England	London	North East	North West	South East	South West	West	Yorkshire and The Humber
Aged 16–17 years	0.231	0.246	0.156	0.286	0.191	0.138	0.211	0.214	0.288	0.204
Aged 18–25 requiring support	0.161	0.126	0.134	0.254	0.091	0.188	0.097	0.152	0.131	0.167
Young parent requiring support	0.187	0.160	0.142	0.174	0.127	0.185	0.135	0.227	0.234	0.312
Care leaver 18–20 years	0.160	0.103	0.233	0.188	0.133	0.064	0.112	0.202	0.151	0.140
Care leaver aged 21+ years	0.221	0.136	0.127	0.253	0.387	0.168	0.236	0.218	0.112	0.231
Physical ill health/ disability	0.141	0.106	0.078	0.182	0.174	0.099	0.149	0.146	0.155	0.105
Mental health problems	0.149	0.112	0.079	0.158	0.184	0.096	0.183	0.122	0.178	0.125
Learning disability	0.206	0.174	0.162	0.168	0.241	0.156	0.190	0.306	0.153	0.244
Sexual abuse/ exploitation	0.240	0.167	0.208	0.237	0.435	0.194	0.214	0.241	0.173	0.226
Domestic Abuse	0.122	0.073	0.082	0.144	0.147	0.097	0.148	0.111	0.118	0.107
Abuse (non-domestic abuse)	0.208	0.206	0.202	0.248	0.235	0.172	0.187	0.184	0.153	0.212
Drug dependency needs	0.178	0.220	0.116	0.165	0.214	0.122	0.139	0.168	0.089	0.190
Alcohol dependancy needs	0.156	0.136	0.138	0.151	0.173	0.134	0.103	0.180	0.094	0.142
Offending history	0.226	0.228	0.132	0.225	0.259	0.190	0.176	0.214	0.181	0.325
Repeat homelessness	0.301	0.286	0.230	0.229	0.352	0.349	0.264	0.363	0.199	0.295
History of rough sleeping	0.269	0.264	0.184	0.249	0.378	0.232	0.203	0.313	0.253	0.316
Former asylum seeker	0.305	0.293	0.191	0.290	0.393	0.161	0.269	0.209	0.078	0.443
Old age	0.124	0.058	0.102	0.165	0.122	0.171	0.077	0.152	0.157	0.086
Served in HM Forces	0.222	0.109	0.163	0.305	0.219	0.209	0.095	0.211	0.192	0.234
Access to EET	0.374	0.464	0.376	0.381	0.338	0.331	0.330	0.424	0.193	0.318

Table 2d Household homelessness ICCs by support needs and region (2020-21)

	All	East Midlands	East of England	London	North East	North West	South East	South West	West	Yorkshire and The Humber
Aged 16-17 years	0.247	0.177	0.217	0.286	0.255	0.260	0.115	0.247	0.172	0.226
Aged 18-25 requiring support	0.167	0.256	0.142	0.150	0.165	0.181	0.078	0.133	0.150	0.254
Young parent requiring support	0.200	0.153	0.192	0.250	0.198	0.288	0.103	0.194	0.074	0.297
Care leaver 18-20 years	0.137	0.197	0.138	0.139	0.141	0.086	0.115	0.143	0.067	0.046
Care leaver aged 21+ years	0.182	0.127	0.168	0.134	0.165	0.132	0.185	0.219	0.142	0.174
Physical ill health/ disability	0.144	0.137	0.095	0.131	0.181	0.145	0.134	0.131	0.147	0.105
Mental health problems	0.149	0.140	0.117	0.105	0.170	0.221	0.117	0.116	0.165	0.131
Learning disability	0.207	0.168	0.187	0.165	0.206	0.231	0.185	0.202	0.215	0.227
Sexual abuse/ exploitation	0.250	0.233	0.214	0.204	0.250	0.303	0.243	0.320	0.117	0.208
Domestic Abuse	0.150	0.088	0.133	0.115	0.096	0.149	0.201	0.137	0.128	0.100
Abuse (non-domestic abuse)	0.231	0.246	0.233	0.219	0.277	0.189	0.203	0.257	0.213	0.211
Drug dependency needs	0.197	0.235	0.150	0.195	0.178	0.150	0.211	0.153	0.126	0.177
Alcohol dependancy needs	0.151	0.187	0.103	0.140	0.129	0.089	0.137	0.188	0.119	0.117
Offending history	0.212	0.224	0.159	0.254	0.238	0.193	0.178	0.212	0.138	0.201
Repeat homelessness	0.283	0.282	0.243	0.251	0.421	0.245	0.235	0.394	0.160	0.333
History of rough sleeping	0.276	0.325	0.257	0.239	0.354	0.344	0.234	0.252	0.177	0.307
Former asylum seeker	0.316	0.376	0.208	0.199	0.322	0.198	0.329	0.303	0.280	0.279
Old age	0.145	0.113	0.160	0.135	0.103	0.126	0.093	0.165	0.221	0.060
Served in HM Forces	0.237	0.146	0.172	0.307	0.228	0.118	0.211	0.204	0.148	0.162
Access to EET	0.351	0.391	0.302	0.366	0.538	0.386	0.223	0.366	0.179	0.567

Results of ICCs by support needs indicate a higher level of heterogeneity among estimates among regions for both years. Values for all categories and regions, for the period of 2021-20, range between 0.05 and 0.57, while for the period of 2022-21, estimated ICCs are slightly lower, ranging between 0.06 and 0.46. This observation also reflects the observed difference in estimated ICCs through time. When comparing estimates between tables 3 and 4, it is evident that values have generally fallen as we move from the previous period to the next. Contrary to the above, the region of London is the most notable exception to this rule, where values have instead mostly risen. Contrary to the variation observed for regions through time, when considering aggregate estimates for all of England (see column 'All'), values seem to remain relatively stable. Higher-than-average ICC values were estimated for the following categories:

- Care leavers
- People at risk of/experienced abuse,
- People with a history of repeat homelessness/rough sleeping
- Former asylum seekers
- People in need of education, employment, or training.

The reasons for the higher observed ICCs for these categories could be related to area-level effects, which these groups may be particularly sensitive to. We would thus expect a larger design effect as a result of the higher estimated ICCs in these cases, and thus larger sample size requirements. As these estimates are specific to England, ICC values of the categories explored may be characterised by some variation in other countries, particularly given the importance of country-specific policies targeted at vulnerable groups experiencing homelessness; conservatism is therefore warranted when considering these values for trials outside of England. However, given the literature's severe lack in availability of ICC estimates for household homelessness and homelessness in general, we believe that the aggregate estimates provided in this paper could serve, at the very least, as a useful starting point for researchers when considering the required power for trials in other countries.

Policing

The tables below show the results of the analysis described

Table 3a LSOAs nested in Constabularies

Street crime type	22-01	22-02	22-03	22-04	22-05	22-06	22-07	22-08	22-09	22-10	22-11	22-12
Anti-social behaviour	0.072	0.058	0.100	0.075	0.071	0.076	0.081	0.067	0.078	0.109	0.073	0.058
Bicycle theft	0.017	0.023	0.020	0.024	0.021	0.019	0.019	0.021	0.011	0.025	0.018	0.028
Burglary	0.032	0.037	0.028	0.031	0.038	0.036	0.036	0.030	0.042	0.038	0.029	0.034
Criminal damage and arson	0.064	0.066	0.071	0.064	0.062	0.056	0.064	0.062	0.067	0.066	0.065	0.063
Drugs	0.030	0.036	0.039	0.029	0.028	0.029	0.032	0.029	0.032	0.023	0.034	0.036
Other crime	0.036	0.028	0.029	0.031	0.040	0.028	0.024	0.039	0.028	0.027	0.026	0.032
Other theft	0.020	0.021	0.019	0.022	0.018	0.013	0.014	0.019	0.018	0.014	0.015	0.018
Possession of weapons	0.030	0.019	0.026	0.015	0.035	0.021	0.024	0.026	0.024	0.024	0.027	0.032
Public order	0.051	0.044	0.058	0.064	0.065	0.070	0.073	0.075	0.068	0.067	0.053	0.056
Robbery	0.012	0.009	0.019	0.011	0.020	0.027	0.020	0.015	0.019	0.018	0.017	0.019
Shoplifting	0.023	0.019	0.017	0.019	0.016	0.021	0.015	0.022	0.017	0.021	0.012	0.018
Theft from the person	0.020	0.012	0.022	0.015	0.017	0.019	0.020	0.021	0.015	0.016	0.025	0.019
Vehicle crime	0.042	0.046	0.045	0.044	0.047	0.043	0.042	0.039	0.044	0.047	0.039	0.045
Violence and sexual offences	0.093	0.093	0.096	0.095	0.100	0.096	0.100	0.093	0.099	0.102	0.088	0.089

Estimates of the intra-cluster correlation rate for all street crimes at LSOA level nested within constabularies suggest weaker homogeneity within clusters. On the other hand, ICCs at LSOA level nested within MSOAs indicate the presence of a stronger clustering effect, with estimates hovering around 0.1-0.2. Values are estimated to be slightly larger for certain crime types such as vehicle crime, theft from person, violence and sexual offences and criminal damage and arson.

Table 3b LSOAs nested in MSOAs

Street crime type	22-01	22-02	22-03	22-04	22-05	22-06	22-07	22-08	22-09	22-10	22-11	22-12
Anti-social behaviour	0.196	0.190	0.310	0.213	0.230	0.211	0.224	0.224	0.214	0.322	0.196	0.187
Bicycle theft	0.101	0.108	0.130	0.143	0.158	0.160	0.154	0.140	0.147	0.153	0.152	0.120
Burglary	0.101	0.125	0.101	0.134	0.106	0.109	0.106	0.108	0.110	0.111	0.098	0.112
Criminal damage and arson	0.157	0.162	0.168	0.161	0.175	0.151	0.157	0.159	0.163	0.155	0.147	0.140
Drugs	0.162	0.152	0.185	0.136	0.134	0.133	0.144	0.168	0.134	0.147	0.136	0.129
Other crime	0.098	0.067	0.110	0.115	0.100	0.111	0.093	0.133	0.110	0.075	0.092	0.078
Other theft	0.128	0.132	0.133	0.121	0.126	0.120	0.134	0.130	0.122	0.152	0.138	0.145
Possession of weapons	0.073	0.066	0.095	0.055	0.050	0.043	0.078	0.052	0.098	0.051	0.098	0.073
Public order	0.142	0.128	0.156	0.161	0.161	0.172	0.185	0.184	0.169	0.163	0.142	0.153
Robbery	0.112	0.112	0.153	0.083	0.158	0.127	0.198	0.143	0.154	0.149	0.118	0.146
Shoplifting	0.094	0.078	0.087	0.087	0.099	0.096	0.078	0.104	0.070	0.049	0.076	0.076
Theft from the person	0.247	0.204	0.243	0.257	0.273	0.269	0.271	0.272	0.218	0.288	0.269	0.269
Vehicle crime	0.153	0.135	0.151	0.133	0.149	0.148	0.147	0.154	0.148	0.166	0.167	0.149
Violence and sexual offences	0.267	0.267	0.274	0.282	0.285	0.280	0.279	0.275	0.275	0.268	0.259	0.258

For outcomes, estimated ICCs by time point and outcome nested within constabularies and MSOAs largely followed the same pattern observed for street crime types, although values were generally higher. There were a few notable outcome types such as the inability to identify and prosecute suspects, which indicated stronger clustering effects, suggesting significant variation between constabularies and MSOAs. Outcome types exhibited more variation in estimated ICCs through time and between outcomes; this observed heterogeneity persisted for yearly ICCs as well.

Table 3c LSOAs nested in Constabularies

Outcome type	22-01	22-02	22-03	22-04	22-05	22-06	22-07	22-08	22-09	22-10	22-11	22-12
Action to be taken by another organisation	0.034	0.048	0.035	0.039	0.064	0.056	0.037	0.034	0.083	0.037	0.045	0.048
Formal action is not in the public interest	0.090	0.107	0.065	0.120	0.063	0.105	0.080	0.046	0.030	0.099	0.087	0.050
Further action is not in the public interest	0.090	0.069	0.065	0.067	0.076	0.052	0.054	0.028	0.097	0.043	0.037	0.046
Further investigation is not in the public interest	0.060	0.111	0.050	0.044	0.043	0.038	0.055	0.048	0.069	0.046	0.050	0.042
Investigation complete; no suspect identified	0.160	0.166	0.177	0.133	0.154	0.150	0.145	0.145	0.179	0.149	0.147	0.146
Local resolution	0.041	0.045	0.037	0.039	0.041	0.042	0.049	0.031	0.045	0.036	0.038	0.029
Offender given a caution	0.143	0.146	0.127	0.154	0.169	0.136	0.150	0.080	0.122	0.109	0.042	0.067
Offender given a drugs possession warning	0.090	0.073	0.188	0.214	0.239	0.019	0.018	0.079	0.108	0.142	0.254	0.144
Offender given penalty notice	0.000	0.161	0.008	0.153	0.013	0.011	0.068	0.021	0.054	0.254	0.009	0.059
Suspect charged	0.090	0.094	0.083	0.097	0.102	0.087	0.095	0.077	0.083	0.084	0.071	0.071
Suspect charged as part of another case	0.124	0.163	0.046	0.153	0.138	0.145	0.227	0.074	0.253	0.083	0.085	0.033
Unable to prosecute suspect	0.090	0.097	0.099	0.114	0.103	0.117	0.122	0.088	0.101	0.066	0.083	0.082

Table 3d LSOAs nested in MSOAs

Outcome type	22-01	22-02	22-03	22-04	22-05	22-06	22-07	22-08	22-09	22-10	22-11	22-12
Action to be taken by another organisation	0.093	0.081	0.073	0.038	0.076	0.106	0.088	0.103	0.093	0.081	0.053	0.049
Formal action is not in the public interest	0.171	0.167	0.059	0.153	0.109	0.112	0.040	0.068	0.092	0.150	0.127	0.037
Further action is not in the public interest	0.141	0.190	0.144	0.117	0.035	0.063	0.095	0.063	0.227	0.140	0.110	0.157
Further investigation is not in the public interest	0.121	0.105	0.183	0.106	0.100	0.114	0.059	0.149	0.104	0.085	0.107	0.086
Investigation complete; no suspect identified	0.434	0.452	0.449	0.285	0.336	0.308	0.343	0.315	0.472	0.307	0.342	0.422
Local resolution	0.111	0.132	0.103	0.129	0.126	0.098	0.089	0.136	0.135	0.150	0.147	0.111
Offender given a caution	0.191	0.155	0.084	0.168	0.183	0.145	0.105	0.094	0.131	0.071	0.103	0.108
Offender given a drugs possession warning	0.107	0.002	0.024	0.213	0.129	0.334	0.270	0.111	0.311	0.243	0.258	0.000
Offender given penalty notice	0.122	0.257	0.076	0.104	0.070	0.000	0.000	0.057	0.054	0.229	0.111	0.060
Suspect charged	0.161	0.155	0.176	0.188	0.169	0.163	0.141	0.169	0.180	0.182	0.173	0.160
Suspect charged as part of another case	0.280	0.421	0.634	0.413	0.159	0.000	0.469	0.401	0.395	0.000	0.179	0.094
Unable to prosecute suspect	0.327	0.318	0.360	0.347	0.342	0.348	0.367	0.343	0.354	0.318	0.322	0.316

These calculations provide a reference list which can be used when designing cluster randomised controlled trials in policing. Estimates are calculated for different clustering levels, providing researchers with a plethora of ICC estimates to be used depending on the characteristics and outcomes of the trial.

Conclusion

In this chapter, we have primarily presented a range of ICCs calculated using publicly available data. These calculations provide a reference list which can be used when designing cluster RCTs in policing, HE participation and homelessness. We note that compared to the ICCs typically assumed in trials elsewhere in education, several of the HE ICCs are very high, particularly for attendance of very selective institutions – detecting a small treatment effect of 0.1 standard deviation change (Cohen, 1988), with statistical power of detection of 80 per cent on propensity to attend Oxford or Cambridge would require a sample of 2180 secondary schools, or roughly two thirds of all secondary schools in England. This suggests that trials might be better suited, from a statistical point of view, to looking at broader definitions of participation. Results from the homelessness data suggest low-to-moderate levels of clustering with ICCs ranging from 0.05 to 0.5. Predictably, certain categories exhibited higher ICC values, particularly for vulnerable groups such as people at risk of/experienced abuse, people with a history of homelessness, and former asylum seekers.

References

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum Associates.

Department for Levelling Up, Housing and Communities (2023) Tables on Homelessness. [online] Accessed at: https://www.gov.uk/government/statistical-data-sets/live-tables-on-homelessness.

EDIT – The Policy Institute (2022) Additional Financial Assistance for People with Experience of Homelessness Trial. [online] Accessed at:
https://osf.io/s8f3t.

Fargo, J.D., Munley, E.A., Byrne, T.H., Montgomery, A.E., and Culhane, D.P. (2013) Community-Level Characteristics Associated with Variation in Rates of Homelessness Among Families and Single Adults, American Journal of Public Health 103(Suppl2) pp.340-347.

Mabhala, M., Esealuka, W.A., Nwufo, A.N., Enyinna, C., Mabhala, C.N., Udechukwu, T., Reid, J., and Yohannes, A. (2020) Homelessness Is Socially Created: Cluster Analysis of Social Determinants of Homelessness (SODH) in North West England in 2020, International Journal of Environmental Research and Public Health18(6) pp.3066.

Sanders, M., Chande, R., Selley, E. and Behavioural Insights Team (2017) Encouraging People into University, London: Department for Education. [Online]. Available at
https://assets.publishing.service.gov.uk/

media/5a82ed3f40f0b62 30269d6cd/Encouraging_people_into_university.pdf (accessed: 14 October 2024).
Sanders, M., Burgess, S., Chande, R., Dilnot, C., Kozman, E. and Macmillan, L. (2018) 'Role Models, Mentoring and University Applications-evidence from a Crossover Randomised Controlled Trial in the United Kingdom', Widening Participation and Lifelong Learning, 20, 4: 57-80.

Sanders, M. and Picker, V. (2023a) Investigating Interventions to Reduce Homelessness Among Care Leavers in Greater Manchester. [online] Accessed at:
https://osf.io/3vcsz.

Sanders, M. and Picker, V. (2023b) Quasi-Experimental Evaluations of Interventions to Reduce Homelessness Among Care Leavers. [online] Accessed at:
https://osf.io/pm65e.

Torgerson, D. (2008) Designing Randomised Trials in Health, Education and the Social Sciences: An Introduction, New York: Springer.

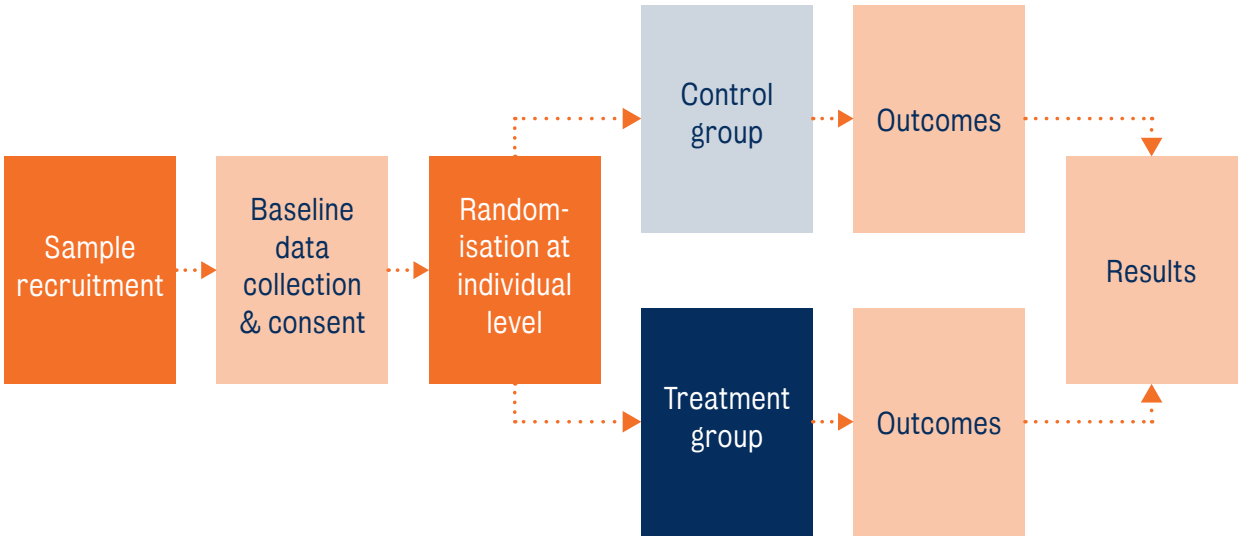
Chapter 12. Complex trials

Introduction

As discussed in previous chapters, the randomised controlled trial (RCT) is, in many ways, a straightforward tool for answering a straightforward question: 'If I do X, what will happen to Y?'. The question of 'what works' is answered through the stacking of lots of these questions together to ask 'what X has the biggest effect on Y?'.

Although the process of running randomised trials can often be anything but simple, the idea, at least, is straightforward. The figure below shows the flow of a standard, parallel designed RCT.

Figure 1 Parallel designs



There are many interventions that we might be interested in, which define testing in such a straightforward manner. For some of these interventions, we may wish to turn to other means of establishing impact – such as quasi-experimental designs.

However, it will often be the case that interventions are complex and there remains a strong reason to wish to conduct a trial. Quasi-experiments may not be possible given the data available or their assumptions; the intervention may be new and untrialled; or we might be interested in the qualitative process evaluation as much as we are the quantitative question of impacts.

Where this happens, we must design trials that can take into account and manage the complexity of an intervention – and we must make a virtue out of this complexity, rather than viewing it as a burden to overcome. The challenges we face are complex, and so perhaps their solutions are too. Importantly, if we only evaluate that which is easy to evaluate, we will jaundice our evidence base in favour of neat, simple solutions.

What is a complex intervention?

There is no formal taxonomy for what constitutes a complex intervention. As such, we attempt to define one here, through a combination of Anders et al (2017) and CEDIL (2022).

Complexity itself is, of course, complex. Hence, there is not one form of complexity. Within the family of complexity, we identify several types of complex intervention:

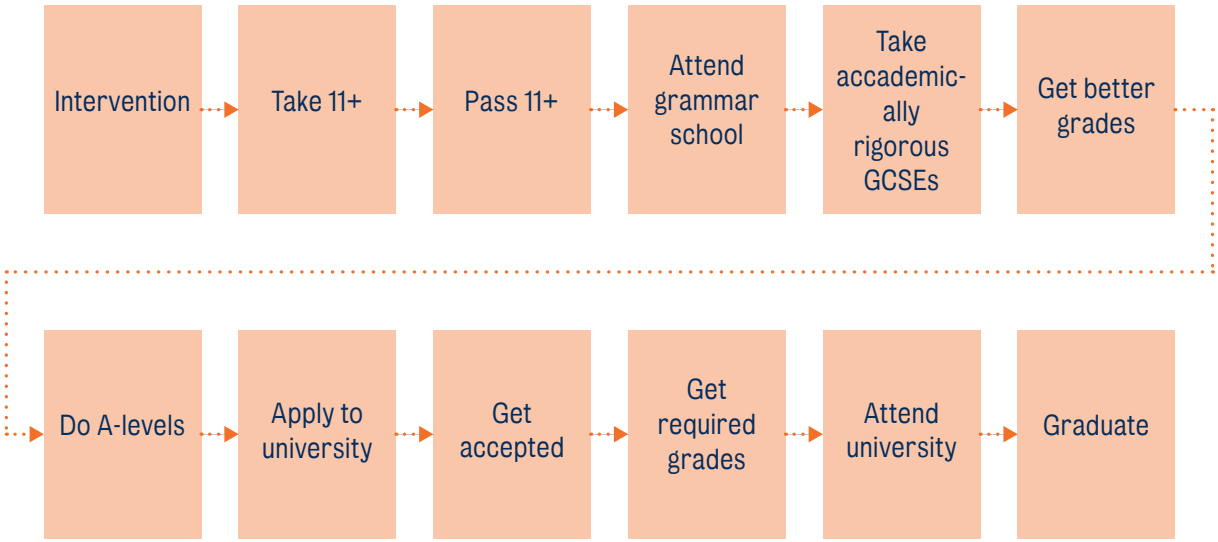
- Long causal chain interventions
- Multi-component interventions
- Multi-target interventions
- System-level interventions
- Evolving interventions

Long causal chain interventions

A long causal chain intervention is one where there are many steps between the input of the intervention and the desired outcome, where achieving the desired ends requires all, or most, of a series of behaviours to take place, or gateways to be passed through.

For example, in higher education access and participation, there is a growing interest in raising attainment well before a young person is due to apply to university – sometimes even as far back as their primary school years. An attainment-raising intervention in a county with academic selection at 11 – that is, one with the 11+ exam and selective grammar schools – might provide tutoring to increase attainment in the 11+, leading some young people to cross the threshold of acceptance to the local grammar school who otherwise wouldn't have done. Those young people would then be more likely to enter particularly academically rigorous GCSEs, be more likely to get good grades in, for example, English, Maths and Science; more likely to take A-levels, and to take the right A-levels for the courses they are interested in, and get the right grades, to receive an offer from a selective higher education provider, and to choose to attend (see the diagram below).

Figure 2 Long causal chains



This is a long causal chain, both in terms of the number of elements, and the amount of time – potentially a decade – between the first step and the intended outcome. Interventions can alhave long causal chains over a short space of time. For example, we have recently conducted a trial, summarised in box 1, which had a relatively long causal chain over a few months.

Testing the effect of bursary information on HE applications

TASO and King's College London collaborated to test whether providing students with information about income contingent bursaries and grants for which they might be eligible is effective at either increasing their likelihood of applying to university or changing the university to which they apply.

In autumn 2022, schools in England were randomly selected to either receive a parcel containing materials for students (treatment schools), or not (control schools). The parcel contained brown envelopes, not addressed to specific students, containing a letter and a booklet. The brochure contained information about income contingent bursaries and grants provided by all universities in England.

The theory of change for this intervention is straightforward. Bursaries and other income contingent grants represent a reduction in the price/increase in the benefit of attending university, and people's choice of whether to attend university and which university they attend is affected by the price.

However, the causal chain has several steps – the school staff must distribute the intervention materials to students in schools, then students must engage with these materials and use them to inform their research on HE options. For any effect to be seen in the HE application data, the students must then make changes to whether and where they apply to HE. Although the time between the intervention and HE applications may be short (just a few months in the case of Year 13 students receiving the materials) there are several steps which must all take place for any impact to take place. And even if there is a change in student behaviour in relation to researching their HE options, it is not guaranteed this will be reflected in HE applications, which is the only outcome which will be observed in this instance.

The effects of a long causal chain by itself is that many, perhaps most, elements of the intervention could be successful, but the final outcome might not be achieved. Alternatively, with long causal chains and long-time frames, participants could be lost to follow up and their outcomes not captured.

Multi-component interventions

Multi-component interventions are those which have multiple discrete elements, which could be separable (and hence evaluable on their own merits), but are brought together as a part of single intervention.

Multi-component interventions are common in practice; for example, the North Yorkshire 'no wrong door' model combines a residential care hub, with:

- a life coach who is a clinical psychologist
- communication support worker who is a speech therapist
- two community hub foster families who are part of the professional team
- community high needs supported lodging hosts for 16 and 17-year-olds, staffed by people who are specially trained and are part of the professional team.

In addition, the 'no wrong door' model also includes a form of intensive family preservation support to prevent young people from entering care – itself an evidence-based intervention.

Evaluating multi-intervention outreach via randomised controlled trials

Multi-intervention outreach and mentoring is a resource-intensive widening participation (WP) activity and requires significant investment of time and effort from higher education (HE) providers and students alike. Programmes usually offer a combination of activities including: mentoring, coaching, information, advice and guidance, campus visits, subject tasters and summer schools, and these activities often engage hundreds of students over a year or more. Multi-intervention outreach is one of the most common approaches used by HE providers and such programmes are associated with positive aspirations and attitudes towards HE (Robinson & Salvestrini, 2020). However, the existing literature provides correlational and contextual evidence on the efficacy of this approach, rather than a causal link between intervention and outcomes for students.

To address these issues,TASO commissioned and oversaw a series of evaluations, partnering with three HE providers (HEPs) to explore the different ways in which multi-intervention outreach programmes could be evaluated, including the use of randomised controlled trials.

For example, historic oversubscription to a post-16 WP multi-intervention outreach programme at one HEP meant not all applicants could be given a place. By restricting access to the most resource-intensive aspects of the programme (a summer school and online mentoring) to a random selection of applicants the HEP had sufficient resources to deliver the remaining aspects of the outreach programme (UCAS application support and study skills activities) to everyone. In essence this was an RCT comparing a business-as-

usual WP programme (relatively high cost) with a lighter-touch version (relatively low cost), helping unpick the effectiveness of different components of the programme.

Long-term outcome data on actual HE entry is not yet available, but interim analysis using UCAS data indicates that in comparison with students on the business-as-usual programme those on the light touch version were no less likely to make an application or firmly accept an offer to study at an HEP, though students on the light touch programme did receive fewer offers. Replication of these results using final outcome data could suggest that university outreach can be spread more thinly to a greater number of students, without compromising on impact.

Multi component interventions can be challenging to evaluate, as different components can be implemented with different degrees of fidelity, and the interventions might intersect and interact with each other. Different elements may also have separate but complementary outcomes, which must be captured using appropriate outcome measures. Interventions that are individually impactful might be more – or less – impactful in concert.

Multi-target interventions

Some interventions aim to support different groups of people in different ways. Many, for example, include an initial assessment of need, and a prioritisation of particular components of the intervention based on different levels, types, or needs. These are multi-target interventions.

For example, an intervention might involve staff from a higher education provider working with a school to identify different groups of students to be supported towards higher education. For some students currently struggling, this might mean additional support with key subjects; for others, whose attainment is on track, it might mean mentoring, with mentors able to suggest additional interventions along the way.

However, the initial assessment, and the collaboration between the school and the HEP is a part of the intervention – and so cannot be separated from the different elements of the intervention – meaning that randomisation, if it is to happen, must be at the school level. As the assessment is not carried out in the control schools, it is not possible to identify the counterfactual participants within those schools for either of the two groups, and so the treated participants across different target groups must be pooled for analysis.

In extremis, multi-target interventions might aim to change different outcomes for their different targets, further complicating the evaluation.

System-level interventions

Some interventions seek to change the entire system in which they operate – this could be at the level of a team, of a locality, a local authority, or a higher education provider – or indeed, a country.

Whole system changes require a combination of changes to processes, changes to structures, and changes to culture. As a result, these interventions might be difficult to tightly define in terms of an intervention theory of change or manual.

Randomisation for these kinds of interventions is particularly challenging, as they are often slow to implement, and the level of randomisation would be very high, necessitating very large, potentially expensive trials. Above a certain level, randomisation may not be possible.

Evolving interventions

Some interventions are designed to evolve over time and to adapt to the circumstances they are being implemented in. This could include mentoring, where the dynamic between mentor and mentee changes over time, idiosyncratically to the needs of the mentee and the link between the two. Similarly, interventions tested through outcomes-based commissioning, are unlikely to remain static throughout the duration of a randomised trial (Anders and Dorsett, 2017).

Interventions may also evolve in response to an evolving context. A particularly dramatic example of this is the need for many interventions in the field to be altered in response to the coronavirus pandemic, and in particular, the national and local lockdowns that it brought about. At a smaller, but no less important level, public service delivery over the last decade has been affected by other changes in circumstances, such as changes in the inspection regime facing children's services, early help, and other areas. Widening participation activities in higher education have similarly been changed by the institution of the regime of access and participation plans.

The impact of the coronavirus pandemic on an evaluation of university summer schools

University summer schools are an on-campus widening participation intervention comprising a range of activities designed to give students an experience of higher education, including a residential stay in student accommodation, subject tasters and social activities. Studies have found a positive correlation with attending a summer school and higher attainment, and application to and acceptance by HE providers (Burgess et al, 2021; HEFCE, 2010; Hoare & Man, 201; TASO, 2021). However, there is lack of causal evidence on the impact on this approach.

To fill this gap, TASO is conducting a RCT of HE summer schools. The design exploits the oversubscription of these interventions, with applicants randomly assigned to the treatment group (receive a place on the summer school) or the control group (do not receive a place). TASO is capturing attitudes towards HE via a survey administered before and after the summer school, however, the primary outcome is enrolment in HE.

Eight universities participated in the trial which evaluates summer schools that took place between June and August 2021. The coronavirus significantly impacted delivery of the summer schools which were designed to be conducted on campus. Due to restrictions preventing pupils from coming onto campus, one university planned to deliver their summer schools in person at two partner schools, but these were cancelled after randomisation due to coronavirus outbreaks. All other university summer schools part of the trial were delivered online and required new design work, differing substantially from face-to-face delivery. Campus tours became virtual and subject tasters were delivered over zoom. Opportunities to socialise were even more difficult to engineer, with one university sending pizza to all summer school participants in order to replicate a group

dinner on screen. In qualitative interviews conducted as part of the implementation and process evaluation, students and staff remarked on the challenge with engagement, the lack of opportunity for more informal conversations with other participants and with student ambassadors, and the 'awkwardness' of being on camera. However, there were also key unexpected benefits to online delivery, such as participants being able to apply for summer schools at non-local universities, and the flexibility of accessing recordings of sessions to watch at a later date.

The RCT continues, as we await long-term outcome data, but the results must be interpreted in light of the significant way the intervention evolved. The project has since been extended to evaluate face-to-face summer schools which took place in 2022.

Evaluating complex interventions

Different types of complex intervention necessitate different forms of evaluation. In the coming pages, we describe different forms of evaluation that might be suitable.

Pragmatic trials

Pragmatic trials are a broad, and not terribly well-defined class of randomised trials. They maintain the rigour that comes with randomisation, but often involve making some kind of compromise to that rigour (particularly around either the manualisation/stability of the intervention or the quality of causal identification) in order to meet the demands of the particular context. Pragmatic trials can be well-suited for evaluating long causal chain interventions, which stipulate a series of dependencies and are embedded in specific, often dynamic contexts. In the face of these nuances, pragmatic trials embed a qualitative process evaluation throughout, which helps to accomplish two things: First, identifying possible outcomes, mechanisms, and subgroups that may moderate effects at each stage of the causal chain at the outset; and second, assessing intervention fidelity along the chain while the trial is underway.

To return to the earlier example of increasing higher education access and participation through an early-stage intervention of targeted tutoring, researchers can use a qualitative process evaluation to identify each causal link comprising the chain between early-stage tutoring and higher education enrolment. This mapping of d, x, x, y, etc. enables the development of stage-specific hypotheses that can be tested individually using conventional quantitative methods. While this exercise could potentially generate a long, unwieldy list of testable hypotheses (which would in turn incur a higher burden in terms of data collection), if certain connections along the causal chain are already well-understood from other studies, researchers can also narrow their focus to estimating effects of less well-understood links (CEDIL, 2022).

Pragmatic trials are implemented widely across a diversity of interventions, and as such, they can be hard to characterise generally. As a guide, Jamal et al. (2015) suggest a three-stage process for pragmatic trial evaluation:

- Elaborate a theory of change and specify the hypotheses to be tested.
- Describe how emerging findings in the process evaluation of the trial will inform the refinement of the hypotheses.
- Test hypotheses using a combination of process and outcome data, paying attention to particular mediators (mechanisms) and moderators identified throughout the process evaluation.

It is worth bearing in mind that given their sensitivity to a trial's context, the external validity, or generalisability, of pragmatic trials can be limited.

Longitudinal trials

Another approach to evaluating long causal chain interventions is through the use of longitudinal data. Cooperation with an ongoing cohort study such as the Avon Longitudinal Study of Parents and Children (ALSPAC) or the Millenium Cohort Study (MCS) could potentially yield high-powered, robust causal effect estimation, especially for outcomes that may take years to come to fruition (such as our example of a primary school tutoring intervention leading to higher education enrolment).

Using long-term administrative data in evaluations of widening participation interventions

TASO is conducting RCTs to understand the impact of multi-intervention outreach programmes and university summer schools (see Box 2 and Box 3 respectively). The primary outcome for both evaluations is enrollment in HE. One HE partner in the multi-intervention outreach evaluation is King's College London (KCL); KCL run the K+ programme designed to support Year 12 students from 'widening participation' backgrounds (including students from lower income families, people who would be the first in their family to attend university, and students of colour) in applying to highly selective universities. The K+ programme was implementedin the 2021-22 academic year and therefore participants will not be eligible to enter HE until September 2023. The universities participating in the summer schools trial also target year 12 students, again not eligible to enter HE until 2023.

In order to assess the impact of both interventions, TASO will be accessing administrative data for trial participants (both treatment and control groups) which will determine whether they have enrolled in HE and which provider they attend (ie, the host university or an alternative provider). This information is accessed through the Higher Education Statistics Agency (HESA). Prior attainment is a key covariate in the trial, as a variable that has an impact on HE enrolment, and this data is accessed via the National Pupil Database (NPD). Subsequently, a matched dataset is required for all trial participants from both HESA and the NPD. The linked HESA-NPD data will be released in 2024. As the data can be accessed for all trial participants, and is not subject to the attrition seen in collecting survey outcomes, it can yield a high-powered, robust causal effect estimation of both summer schools and multi-intervention outreach programmes.

A cohort study works by identifying a large sample of research participants that share a common characteristic, such as a specific birth month and year, and then following up with participants at regular intervals to gather specific survey data, referred to as panels. Cohort studies are usually thoughtfully designed, with significant attention paid to key measurement questions, consistency over time, and mitigating participant attrition. They often publish cohort data alongside extensive how-to documentation for researchers, including weighting suggestions where appropriate. All this, plus large sample sizes and the consistency of follow ups, make cohort studies very appealing sources of data for intervention trials.

How can a trial be embedded in a cohort study in principle? While cohort studies can and are used with quasi-experimental methods to identify, for example, the effects of parental smoking on child health outcomes (a treatment that would be ethically and practically impossible to randomise), embedding an RCT into a cohort study naturally requires more logistical overhead and coordination across stakeholders. Researchers could collaborate with a cohort study to select a randomised treatment group within the cohort and introduce an intervention, then over time, data are collected on this treatment group and the larger cohort to see if and how outcomes differ over time. To return to our tutoring and higher education example, families of treated children could receive a voucher or cash transfer with the goal of enrolment in tutoring. Naturally, this is a relatively expensive encouragement design, but a lighter intervention could simply be an information treatment for parents and caregivers on the value of tutoring for long term academic success. After the intervention, follow up panels with the treatment and control groups would likely require additional questions to capture expected outcomes over time.

Embedding nudge interventions in a cohort study to study the effect on HE participation

King's College London, University College London and TASO are collaborating on a trial to test whether light-touch, low-cost 'nudge' interventions can help widen participation in HE. The intervention in this trial is a combination of approaches that have previously been shown to impact on higher education application and participation. This includes:

- Letters targeted to the individual from role models who are existing students from similar backgrounds.
- Text messages emphasising the financial benefits of higher education participation.
- Text messages emphasising the financial support available to lower income students and providing links to resources.
- Text messages emphasising the opportunities for belonging at a higher education institution.

Participants for this trial are drawn from the COSMO cohort study, which is a national longitudinal study examining the impact of the coronavirus pandemic using a representative sample of over 13,000 young people.

During the second wave of this study, Year 13 students were randomly allocated to either receive the intervention or not as part of a simple two-armed trial. These students will be tracked over time and long-term education outcomes will be collected as part of the COSMO study. These outcomes will also then be analysed as part of the embedded RCT to understand the effect of the interventions on HE applications.

Benefits of using cohort data

There are both methodological and logistical benefits of this coordinated approach with cohort studies. For one, researchers can make a strong case for baseline comparability between treatment and control units–children will be the same age, and much would already be known about individuals within each group such as socio demographic details and health status. Cohort studies are also usually 'large N' studies with significant efforts to minimise attrition between waves, and provided that a treatment can be implemented with a large enough group, the analysis can be well-powered enough to estimate even modest effect sizes (Sanders & Stockdale, 2023). Further, the additional panel data collected on individuals enable subgroup analysis, to capture possible differences in treatment effects between ethnic groups, incomes, etc.

Logistical benefits include ease of following up with and collecting data from study participants. Participants benefit as well from this logistical efficiency, since the addition of a few additional questions within the larger panel is a relatively small burden. An important caveat to point out however is some cohort studies charge significant fees for adding questions to their panel, so while cohort studies may aid in easing the burden of independent data collection, costs must be considered as well (Sanders & Stockdale, 2023).

Drawbacks

Naturally any cohort study should be scoped first for relevance: the data collected and other details like the pace of follow-up intervals may not be appropriate for a given intervention trial. Beyond this, coordination with cohort studies can introduce a degree of logistical overhead that may make collaboration costly: any RCT would need to meet the particular ethics requirements the cohort study is party to, the funders of the cohort study would likely need to be consulted and permissions sought, and GDPR compliance would need to be achieved through specific data protection processes, to name a few high level considerations (Sanders & Stockdale, 2023).

Additional considerations and trade-offs

All these considerations and costs mean that the most appropriate interventions to test with cohort studies would likely be those that have already shown some promise, rather than completely novel interventions (Sanders & Stockdale, 2023). The purpose of the trial then is to better understand the long-term effects of a particular intervention, rather than discern whether the intervention has any effects at all. Further, stakeholders within the cohort studies and the research group may differ on whether interventions should be more light touch vs. more intensive. Light-touch interventions (for example, our information treatment) benefit from being low cost to implement and minimising possible diversion of the treatment group from the rest of the cohort, which may risk the overall integrity of the cohort study. However, light touch interventions are less likely to create discernible effects long term, obviating a lot of the value of working with cohort data to begin with. More intensive interventions (such as cash transfers) are more likely to see long term effects, but as they are more costly and present greater risk to diverting part of the cohort, treatment groups will likely be smaller; this potentially introduces greater uncertainty in estimating effects later.

Complex trials

As discussed above, multi-component complex interventions combine multiple intervention activities that interact with one another within a context and aim to produce change. An example of multi-component intervention could be the piloting of a new academic program in schools that embeds reading instruction across the curriculum, with an eye toward closing reading skills gaps among low-income pupils. While the intervention could simply be conceived as the new curriculum, there are many components, or 'active ingredients,' (Oakley et al, 2006) that contribute to the success of the pilot, including supportive school leadership, availability of professional development and ongoing support for teachers, the existence or establishment of interdepartmental collaboration spaces, norms around lesson planning and sharing, and responsive formative assessment strategies, to name a few. What makes the curricular intervention 'complex' is the expectation that each ingredient of the intervention will interact with one another to create a kind of recursive causality, where making improvements in one area may influence the effectiveness of other ingredients, and vice versa. Achieving predicted 'tipping points' could produce a virtuous cycle of improvement, where the initial 'cause' is hard to disentangle (Anders et al, 2017).

There are numerous possible approaches to evaluating a multi-component complex intervention, with differing levels of flexibility for given constraints. Broadly, there are three approaches: implementing a randomised control trial (RCT; the most desirable but also most restrictive/least flexible to accommodate constraints), Quasi-experimental designs (QED) that exploit a detail of the intervention such as timing or geography to create plausible treatment and control groups, and where neither is possible, a non-experimental evaluation that could inform RCTs and QEDs in future evaluations of similar programs.

RCTs for complex interventions

While an RCT is generally regarded as the most robust evaluation option for estimating effect size, the restrictive nature of RCTs may make them impractical to implement with complex interventions. Three considerations that may rule out undertaking an RCT include organisational overhead costs during recruitment, achieving adequate sample size to sufficiently power the RCT, and lengthy outcome timelines.

To return to the reading intervention example, schools would need to be recruited that could plausibly launch the intervention, which means assessing school capacity and readiness for implementing the intervention, followed by schools being randomly allocated to treatment and control groups. This organisational overhead-establishing a relationship between the school and research team, gathering relevant information, generating buy-in and commitments-comes at a cost for the school, and risks creating an incentive for a control-allocated school to implement the program anyway, and biassing results.

Achieving adequate sample size to power the trial can compound these costs to schools, especially if the predicted timelines for detecting an effect are lengthy. If the predicted effect size is relatively modest, more schools will need to be recruited in order to detect these effects, which generates more cost. If the predicted effect is expected to take months or years to come about, this again creates more costs and risks attrition of control schools and the potential for other interventions to confound the outcomes.

Where RCTs are not possible or desirable, QEDs can still generate compelling evidence while managing resource constraints. While QEDs are often known for so-called 'natural experiments,' where existing data is opportunistically analysed post hoc and features of the intervention are exploited to create plausible control and treatment groups, QEDs can also be employed at the outset of a trial to mitigate some of the costs associated with RCTs, such as random allocation of treatment and generating a large enough sample. Here we will discuss matching and difference in differences (DiD) as QED approaches that could be used to evaluate our reading intervention, though others certainly exist.

The basic principle of matching is easy enough to understand.For every treated unit, find a control unit that 'matches' the treated unit in key characteristics, then compare their outcomes. The idea is that the control unit demonstrates what would have happened with the treated unit, absent the treatment. The challenge with matching is selecting relevant characteristics and then choosing a fitting matching technique. For the former, a researcher could draw from existing databases such as the national pupil database to match schools on proportion of pupils that are eligible for pupil premium, with special education needs, have English as an additional language (EAL), have a minority background, etc. Naturally the more characteristics are matched, the more challenging it is to find plausible matching schools, but if too few or the wrong characteristics are selected, this threatens the validity of the comparison. Once characteristics are selected, one of several matching techniques (eg nearest neighbour, calliper, kernel, exact) can be utilised to create a control group of schools that can then be analysed alongside the treated schools to estimate the effect size. Since both of these decisions-characteristics to match and matching technique-are somewhat arbitrary, grounded in the wisdom of the researchers and the constraints of available data, it is a good practice to conduct at least five robustness checks, wherein alternative (but plausible) specifications are used to re-run the analysis and compare estimated effects.

Difference in differences offers something of a way out from the arbitrariness of selecting relevant, observed characteristics. Instead, DiD uses a longitudinal data approach, and assumes that, observed over the same time period, controlled and treated units would demonstrate similar trends (known as the parallel trends assumption). If treated units show deviation from previous trends after treatment, researchers can make a case that the change is attributable to the treatment. Unlike matching, treated and controlled units do not necessarily need to resemble each other at baseline, but similar to matching, DiD requires access to unit-based data prior to the intervention. Ideally, in order to establish credible parallel trends, researchers would have access to at least two measurements of the relevant outcome variable (for example, scores on standardised reading assessments) prior to treatment. One of the risks with a DiD approach is the possibility that other programs or interventions coincide with the trial and confound the relationship between the intervention and outcomes (naturally this risk grows the longer a trial lasts).

Finally, matching and DiD can be combined to enhance the credibility of estimated effects and reduce the burden of choosing a matching protocol. Instead of matching by characteristics, units are matched on trends, ie treated and controlled units are changing (or not) in similar directions and rates of change. Then DiD can be used to estimate if and how the treated, matched units deviate from their assumed path, absent of treatment.

Multi-stage trial protocols

Whether employing an RCT or QED approach, a key asset for a complex trial is a multi-stage trial protocol, which serves as a form of pre-registration that is flexible to the demands of a complex trial. Pre-registrations, wherein researchers specify methods, relevant data, and testable hypotheses–typically are written at the outset of a trial, but in the case of a complex trial, mechanisms and context-specific details such as implementation fidelity might not be known at the outset. A multi-stage trial protocol devolves the pre-registration process into three components: an evaluation protocol, an implementation and process evaluation (IPE), and finally hypotheses formation based on the findings of the IPE.

First, the evaluation protocol is specified at the very beginning of a trial, and as with any pre-registration, researchers should strive to follow the evaluation protocol as closely as possible. An evaluation protocol would include analysis methods, details of the intervention, sampling process, and proposed outcomes, as well as indicating timelines on when further stages of the multi-stage trial protocol will take place.

At the end of the experimental period, but ideally before evaluative data are made available to the research team, the implementation and process evaluation provides qualitative insights about intervention fidelity and important contextual details that may suggest testable mechanisms during the data analysis stage.

Finally, informed by findings in the IPE, researchers can specify and test hypotheses using the data gathered. Taken together, these three stages should be written up and published as the second-stage protocol, building on the original evaluation protocol.

Adaptive trials

One of the argued strengths of randomised trials conducted well comes from their rigour and in part from their rigidity. Specifically, the ability to pre-specify how a trial will be conducted, and crucially what analysis will be conducted and how, through the publication of a protocol in advance.

The approach of publishing detailed protocols, which specify trials down to individual regression models to be used, has the benefit of tying the hands of researchers and evaluators. In the absence of these restrictions, it would be possible for researchers to make analytical decisions that favour finding spurious a statistically significant effects of the intervention, by conducting many analyses and choosing to report those that produce positive findings – what is known as Hypothesising After Results are Known (HARKing) (Kerry, 1998).

This rigidity of randomised trials helps to ensure that the research conducted through them is credible. However, it produces challenges when we are considering the evaluation of complex interventions. Specifically, where the intervention is complex, and has several hypothesised potential impacts or causal routes to impact, specifying analysis up front prevents us from learning during the trial.

Resolving this tension between flexibility and rigidity requires us to identify more precisely what we expect to gain from each of the two.

Strengths of rigidity

There are two strengths granted by rigid, rigorous protocolisation. The first is that it allows the trial itself to be replicated, and for readers to understand how the trial was intended to be conducted, and what the intervention is to a high degree of specificity. The second is that it prevents HARKing, through pre-specification of analysis.

These two benefits are separable in terms of when they need to occur. The first benefit must be attained before the trial begins – it must describe the shape of what is to be done during the trial period. The benefit of statistical pre-specification, however, can be attained later, as long as analytical specifications are agreed and published prior to the analysis taking place, and ideally before the final endline data are received.

Strengths of flexibility

The main benefit of flexibility is that it allows us to learn through the process of the trial, and to generate hypotheses in response to the empirical reality of the trial happening on the ground.

This benefit cannot occur before the trial begins – but it could, in many cases, be achieved prior to the endline data collection for the trial.

Reconciliation of strengths

These strengths can be brought together in trials which deviate only slightly from the canonical approach to trials.

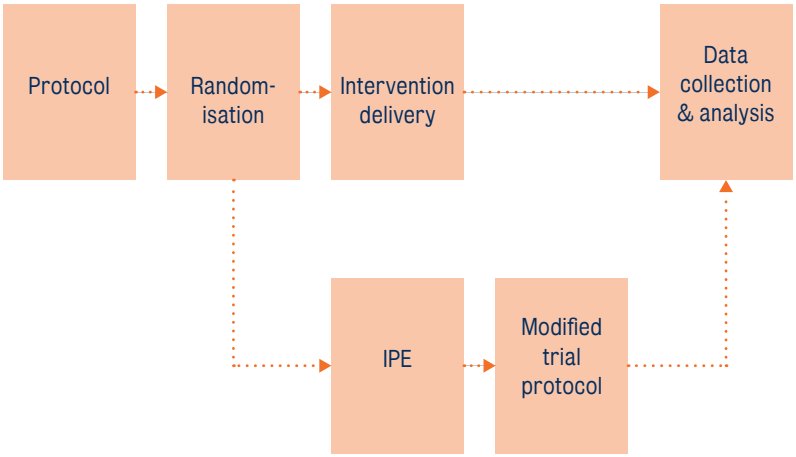
It is a requirement that a trial protocol is produced and published ahead of time, which details how the trial itself will take place, and plans for data collection. This is how things are currently done. However, in an adaptive trial, we can either not publish a statistical analysis plan, or we can publish one conditional on later findings.

Over the course of the trial, various other forms of research, mostly in the form of implementation and process evaluation, can be conducted, which will give insights into how the intervention is conducted; which of several outcomes at the most likely to be effected, and which of many subgroups are most likely to experience benefit from the intervention, as well as picking up issues of fidelity and any unintended consequences.

On the basis of the findings of these components of the evaluation, hypotheses can be developed for statistical testing. These might necessitate the collection of more or different data at the endline than had been anticipated, so that these new hypotheses can be tested. All of this can be combined into an initial findings report for the IPE and a statistical analysis plan which can be published in advance of endline data collection.

Taking this approach has the dual benefits of maintaining the rigour of the trial through pre-specification, while allowing us to capture the complexity of the intervention and its effects through flexibility. The approach is shown in the diagram below.

Figure 4 Adaptive trials



This kind of approach is likely to be most attractive when;

- Trials are long and interventions effects are likely to emerge over time.
- The intervention is a complex and/or whole system evaluation where different groups may differentially benefit.

Factorial trials

When we have an intervention with a large number of component parts, a factorial trial may be a strong option.

A factorial trial tests different interventions in various different combinations, in order to isolate both the individual contribution of each intervention component, and (sometimes but not always) the interaction effects between them.

We can consider the simplest case of this kind of trial, where an intervention has two 'active ingredients' – let us say that these are tuition and mentoring – which can be separated or delivered together.

A factorial design gives us four different possible combinations of interventions to be tested; a control group with no intervention; a group that receives tutoring, a group that receives mentoring and a group that receives both. We can present this in a number of ways – either as a grid, or as a diagram, both of which can be seen below.

Table 1 Factorial trial design

	Control	Tutoring
Control	Control, Control	Control, Tutoring
Mentoring	Mentoring, Control	Mentoring, Tutoring

A design with effectively a four-arm trial, with participants assigned at random to one of the four different cells on the grid, seems as though it gives us the ability to test the active ingredients of the intervention to see if they are making a difference together or in isolation.

However, this approach is not without its problems. Typically, we power our trials to detect effects of a reasonable size with 80 per cent probability, relative to a control group that do not receive the intervention. In this case, this level of statistical power is designed to detect the difference between any one arm and the control – but there are two main obstacles.

Multiple Comparisons

Test statistics are designed to give a level of confidence in the directionality of an effect, given the properties of the data. The more tests that we run, the higher the probability of one of these tests giving false positive. It is therefore necessary to adjust our test statistics in order to account for the fact that we are conducting multiple tests. A common approach used is the Bonferroni method, which is aggressive in terms of its sample size implications, but which you may want to adopt as a straightforward tool when designing your study to ensure that you have sufficient power.

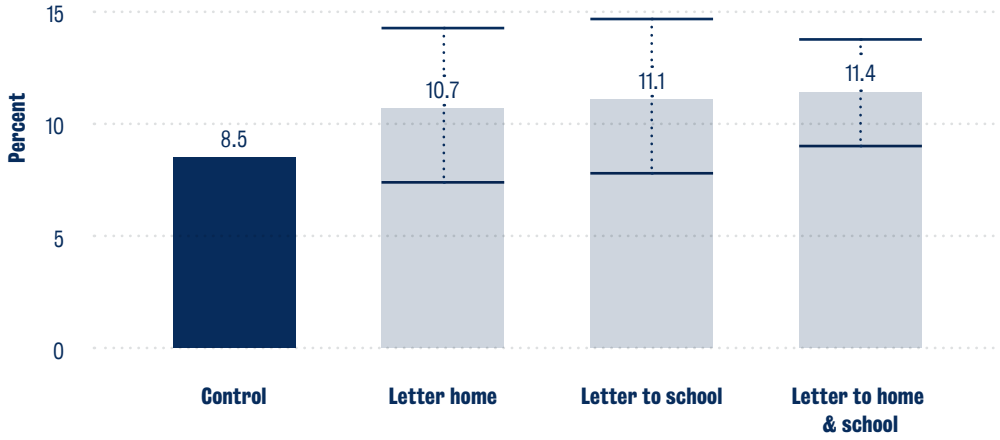
If we consider the four armed trial, and assume only comparisons between the different arms and the control condition, we have 3 tests instead of 1, and so using the Bonferroni approach, our new p value of interest is 0.05/3 = 0.0166. In a simple individually randomised trial, with 80 per cent power aiming to detect an effect of 0.2 standard deviations, for a two armed trial we would need 393 participants per arm, or 786 participants in total. For a four armed trial adjusting for multiple comparisons, we'd need 525 participants per arm, or 2100 participants in total – of 2.6 times as many as we'd need for the two armed trial, or roughly a third more than we'd need if we were to have four arms and no correction for multiple comparisons. This change makes recruitment of participants harder and makes the trial more expensive.

Powering for interactions

In a factorial design, we may not simply be interested in whether or not the interventions and their combinations outperform the control group, but whether they outperform each other. In particular, you might be interested in whether the intervention in combination is greater than the sum of its parts.

To take an example, in the 4 armed trial conducted by Sanders et al (2023), which used different versions of letters from a role model to try and increase attendance at selective universities. In the study, participants are assigned to receive no letter; a letter sent to their school; a letter sent to their home; or letters sent to both places, with the outcome being the proportion of recipients getting accepted to selective universities. As we can see from the figure below, all of the interventions outperform the control group, but only the combined treatment group has a significant effect compared with the control. Hence, we can say that two letters performs better than no letter. However, the difference between the combined treatment and either of the other treatment conditions is not statistically significant. As such, we can say that two letters is better than no letter – but we can say neither that one letter is better than no letter, nor that two letters is better than one letter.

Figure 5 Single and combined intervention effect sizes



Given that a factorial design in this case is intended to identify these kinds of differential impacts, we need to ensure that our tests are well powered to detect effects between groups. The Statistician Andrew Gelman recommends that in order to be confident of detecting these kinds of interaction effects, we should recruit samples that are 16 times the size as if we were just testing comparisons against the controls – but the exact number will depend on how powerful you anticipate the interactions to be, and so it is best to conduct your own calculations.

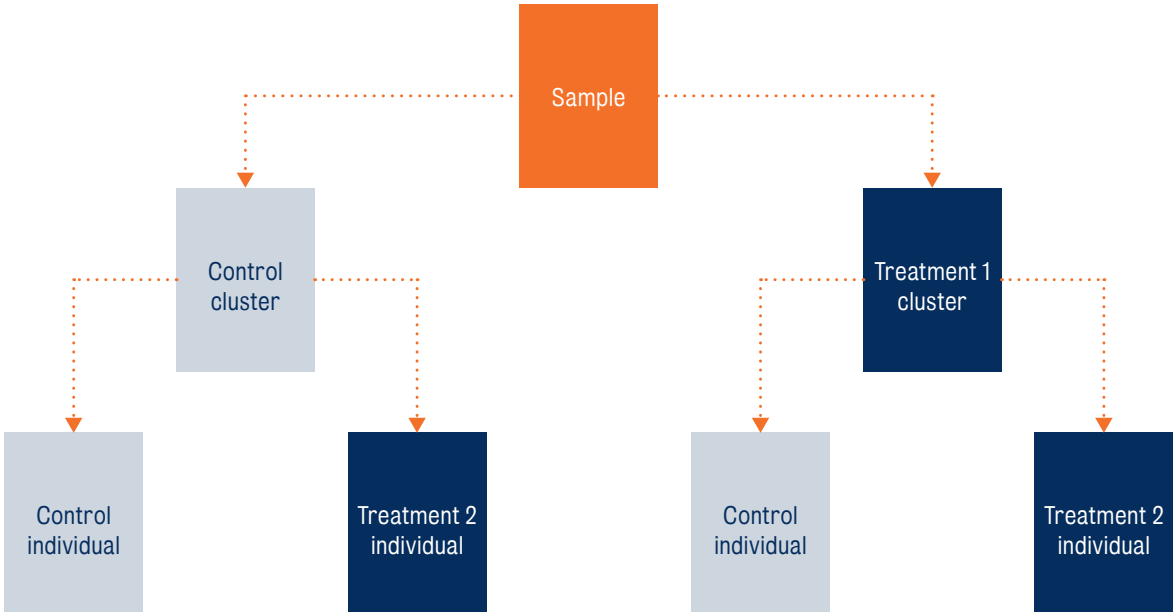
Split plot trials

Split plot trials are a special case of factorial trials, where it is possible to test factorial designs with the same kind of grid structure as shown above, but in a way which is more statistically efficient than the examples given above.

For a split plot to be viable, you need interventions that can be randomised at different levels to each other. To take a simple example, we could consider two interventions that are to be delivered in a school setting, but where one could be randomised at the level of a class or form group, while the second must be randomised at the level of the school itself. For example, this might include the visible classroom intervention, which can be delivered to individual teachers (and their students), and the embedding formative assessment intervention, which involves whole-school changes in practice.

Because these interventions could be randomised at different levels, we can do so, creating a trial diagram like that below – where different levels of the vertical axis relate to different levels of randomisation.

Figure 6 Split plot trials



The split plot design has a number of advantages which can be summarised as;

- Units like schools, employers, higher education providers, and so on are guaranteed that they will receive some of at least one of the interventions, and so are likely to be more engaged.

- Compared with a cluster randomised trial with four arms randomised at the same level, a split plot requires less sample, because it (a) reduces the size of the clusters, and (b) has clustering at a lower level. In extreme cases, introducing the two additional arms with lower levels of randomisation might require close to no additional sample if the intra-cluster-correlation rate is high.

Stepped-wedge trials

Where interventions are large and complex – for example interventions which must be delivered at a whole system level – and in particular where the resource for delivering the intervention is both high and scarce, limiting the rate at which it can be rolled out to many units all at once, a stepped-wedge trial as discussed in Chapter 4 might be a good option.

This approach is not without its challenges – the delivery of a complex intervention, into complex systems, in a random order, means that delays are perhaps inevitable, and there will be a desire to reorganise the rollout to respond to evolving circumstances on the ground. Stepped-wedge trials also require data collection at the end of every 'step', which can be burdensome of some organisations.

Stepped-wedge trials for complex interventions are therefore most likely to be appropriate when;

- The intervention can only be delivered to less than half of the sample at any one given point in time (usually quite a bit less than half).
- The intervention is anticipated to have short or medium term effects, so having some units receiving the 'treatment' for longer than others is useful. Long term effects cannot be captured as all units are treated by the end of the trial.
- Data collection is using administrative records, minimising data collection burden each step.

Conclusions

In this chapter we have considered different types of complex intervention, and how they can be evaluated using a number of different types of design. The exact approach that is optimal under any given circumstance will of course depend on the context, and will require the work of skilled evaluators. Nonetheless, we hope that this guide proves useful.

References

Anders, J. D., Brown, C., Ehren, M., Greany, T., Nelson, R., Heal, J., ... & Allen, R. (2017). Evaluation of complex whole-school interventions: Methodological and practical considerations.

Anders, J. D., & Dorsett, R. (2017). HMP Peterborough Social Impact Bond-cohort 2 and final cohort impact evaluation.

Burgess, A. P., Horton, M. S. & Moores, E. (2021). Optimising the impact of a multi-intervention outreach programme on progression to higher education: recommendations for future practice and research.

HEFCE. (2010). Aimhigher summer schools: Participants and progression to higher education. Higher Education Funding Council for England.

Hoare, T., & Mann, R. (2011). The impact of the Sutton Trust's Summer Schools on subsequent higher education participation: a report to the Sutton Trust. Bristol: University of Bristol, Widening Participation Research Cluster.

Jamal, F., Fletcher, A., Shackleton, N., Elbourne, D., Viner, R., & Bonell, C. (2015). The three stages of building and testing mid-level theories in a realist RCT: a theoretical and methodological case-example. Trials, 16(1), 1-10.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. Personality and social psychology review, 2(3), 196-217.

Robinson, D., & Salvestrini, V. (2020). The impact of interventions for widening access to higher education: A review of the evidence. Education Policy Institute. https://epi. org. uk/publicationsand-research/impact-of-interventions-for-widening-access-to-he.

Sanders, M., Chande, R., Kozman, E., & Leunig, T. (2023). Can Role Models Help Encourage Young People to Apply to (Selective) Universities: Evidence from a Large Scale English Field Experiment. Widening Participation and Lifelong Learning

Sanders and Stockdale (2023) What Works and cohort studies. https://www.kcl.ac.uk/policy-institute/assets/what-works-and-cohort-studies.pdf

TASO (2021). An investigation into the relationship between outreach participation and Key Stage 4 (KS4) attainment/HE progression.

Oakley, A., Strange, V., Bonell, C., Allen, E., & Stephenson, J. (2006). Process evaluation in randomised controlled trials of complex interventions. Bmj, 332(7538), 413-416.

Chapter 13. Split plot trials

Introduction – What is a split plot trial?

There are cases where we want to know whether multiple interventions–when combined together–can have a differential effect compared to the interventions taken on their own. In plain terms, can interventions taken together have an impact that's greater than the sum of the parts (or conversely, can interventions act against each other and attenuate effects)? Split plot trials are one way to help answer these questions, and are especially useful in cases where randomisation is restricted in some way. Split plot trials are unusual, in that unlike the more common flat -plot trials, randomisation occurs at multiple levels, with one intervention being assigned at the individual level, while another is randomised at the level of a cluster.

Split plot trials have long enjoyed popularity in agricultural sciences and industry (Yates, 1933; Parker et al., 2006), but only recently have started to appear in social science. The two main circumstances that make a split plot trial most appealing are: first, the desire to test two or more interventions in a way that captures the individual effect of each intervention and the combined effect; and second, at least one of the interventions is infeasible to allocate on an individual basis, but can be allocated to clusters (say, at the level of a classroom or local authority).

The design of a split plot trial is intuitively easy to understand, but they can get increasingly complex as more interventions are tested, so let's start with the simplest version. Let's say we have two interventions, A and B. We want to evaluate the interventions in combination with each other, but also in isolation, while also preserving a group that receives neither intervention. Intervention A is the 'hard to change' factor (ie a treatment that, by its nature, is difficult to allocate on a granular level, such as the individual level or a 'lower level' cluster), so we start by randomly choosing clusters across the whole sample to receive intervention A, ie the 'whole-plot' intervention. Within each of our clusters (both those receiving intervention A and those not receiving intervention A), we randomise intervention B (the 'split-plot' intervention, creating treatment and control groups for B, nested within the clusters that are randomly allocated to receive (or not receive) intervention A. We now have a 2x2 treatment allocation, resulting in four arms:

Arm I: A = 1, B = 1

Arm II: A = 1, B = 0

Arm III: A = 0, B = 1

Arm IV: A = 0, B = 0

We have a treatment arm that receives both interventions, two treatment arms that receive either intervention A or B, and a control group that receives neither. This approach can be scaled up to include a third intervention C (which would double the arms), or test treatments with three or more arms, etc.

As mentioned, there are a couple reasons that make split plot designs appealing. First, often we are interested in fitting second-order models and seeing interaction effects between interventions, and second, sometimes individual randomisation of both interventions complete randomisation is not feasible – treatments can have spillover effects within groups, or the treatment allocation itself requires a group format (say a classroom intervention), or resources to allocate treatment are limited, making it cost-effective to deliver at a higher level (Cortes et al., 2018). This second reason differentiates split plot designs from randomised complete block design (RCBD). While RCBD also randomises within blocks or groups, the idea behind RCBD is to try to ensure balanced treatment allocation across relevant groups. In split plot designs, the hard-to-change intervention is deliberately chosen as the 'block' for feasibility's sake, and the easier-to-change intervention B applied to the nested units (Altman & Krzywinski, 2015).

Split plot designs have a couple benefits that distinguish them from a traditional individually randomised design. The first is cost and feasibility: some interventions are cost prohibitive or otherwise impractical to randomise on an individual level, but are achievable at a higher order group (or 'whole-plot') level. A classic higher order intervention taken from agriculture is irrigation design: it is not practical to implement a different style of irrigation at random by, say, field row; it is much easier to randomise irrigation at a plot or field level, and then randomise another treatment at the 'split-plot' level (this is actually where we get the name 'split plot'; it is a term borrowed from agricultural trials).

The second main benefit is that split plot trials can improve statistical efficiency, and by this, we are referring to the attenuation of the error term. By assigning interventions at two different levels–group, and then individual–we create more homogenous groupings or blocks, within which there will likely be less variation in outcomes than if we had randomly assigned both interventions at the individual level. This approach potentially yields more precise effect estimates with smaller standard errors for the individually randomised, or split-plot intervention, which can be valuable for finding statistically significant results in cases of interventions with small to medium effect sizes.

Enhancing the statistical efficiency for the split-plot intervention comes with a trade-off for the higher whole-plot intervention, however. Since there are necessarily fewer independent units of analysis available at this level, the evaluation may be under-powered for this intervention, which means less certainty in the effect size estimate and a higher error term (for more detail on quantifying this correlation, see chapters 2, 6, and 11). For this reason, it makes sense to test the intervention we are most interested in at the split-plot level, wherever possible; we can gather more precise estimates at this level, but at the expense of the whole-plot level. Alternatively, it might make sense to test lower cost, lower impact interventions at the split- plot level, as we anticipate that identifying effects of these interventions requires more statistical power compared with a relatively more impactful intervention.

One of the key benefits of a split plot trial is the ability to report the interaction effect between two or more interventions (ie the study arm whose treatment status equals 1 for all interventions). However a related drawback of the split plot trial design is the added complexity in both analysis and reporting of effect sizes. While interaction effects themselves are fairly easy to calculate, explain, and understand, split plot trials are complicated by having multiple error terms, one per level. In a traditional completely randomised design, there is only one error term to manage, making reporting of standard errors and statistical significance for each effect size straightforward. In the case of split plot trials, effect sizes for the whole-plot factor and the interaction must incorporate this extra error term, introducing difficult-to-quantify uncertainty to the interpretation of split plot trial results. We will return to these challenges and peculiarities in the analysis section, following the example below.

Split plot trial example: Improving college attendance through a two-part intervention

A split plot trial might be used in colleges to assign people to more than one treatment, where the natural level of randomisation might vary between different interventions. For example, we might be trying to improve attendance at school with two interventions – one of which is an incentive scheme whereby a particular level of attendance (for example, more than 95 per cent), is rewarded with either money or a classroom activity (similar to Burgess et al, 2021), and another in which students' parents are sent text messages informing them of their child's attendance each week and how that compares to their peers (similar to Chande et al, 2017).

As evaluators, we might decide that the incentivisation scheme cannot be randomised at the level of the individual student – because they are extremely likely to tell their peers about it – and so opt for randomisation at the level of the tutor group. Texts to parents, however, are more easily randomised at the level of the individual student, as there is less likely to be conversation among students about this intervention. We can then design a split plot trial that looks like figure 6 on page 139; with each class randomly assigned either to be treated with the incentives scheme or not to be, and each individual within each class randomly assigned to either receive the parental texting intervention or not to.

We could also have a split plot with three levels – for example if we also wanted to test loss aversion financial incentives, in which teachers are given a bonus at the start of the school year which they lose if their students don't achieve their targeted trades (as in Fryer et al, 2022), this could be randomised at the level of the school (as giving different incentives to different teachers in the same school is likely to be impractical), giving three levels of randomisation.

Example of split plot trial

An example of a split plot trial is one carried out by the Behavioural Insights Team as a part of their Behavioural Research Centre on Adult Skills and Knowledge (ASK). This trial sought to test three interventions, developed by different academic teams. This included a Grit Intervention developed by the University of Pennsylvania's Angela Duckworth; Values Affirmation, developed by Geoff Cohen from Stanford University, and a Study Supporter Intervention, developed by Harvard's Todd Rogers. The three interventions all had the same outcomes in mind — increasing attainment of people resitting their GCSEs — and took place in the same settings (Further Education Colleges). However, the number of further education colleges in England is fairly small (around 280, roughly one tenth the number of secondary schools), meaning that sample size was at a premium. It was determined that while Grit and Values Affirmation interventions, which were completed

in class time, would need to be randomised at the level of the class within a college, the Study Supporter intervention could be randomised at the individual level. Therefore, to maximise power, the study was designed as a split plot. First, colleges would be recruited to the trial. Then, classes within the college would either be allocated to the control condition; to Values Affirmation; or to Grit. All students within those classes would receive that intervention (or none, in the case of the control). Students within all of those classes were then randomised, individually, either to receive the study supporter intervention, or the control condition. This design maximises statistical power for a given sample size (compared to randomising classes to four groups) and also allows for analysis of the interaction effects between Study Supporter, and Grit & Values Affirmation. Calculating statistical power

Calculating statistical power

As with a traditional randomised trial, split plot trials require power analyses to establish a viable sample size before the trial begins. We cover the necessary steps to conduct a power analysis in detail in the Sample size and power chapter, so we will keep this section brief and confined to relevant considerations for split plot trials.

The two distinguishing factors of split plot trials–the testing of two or more interventions in combination, and the random allocation of the hard-to-change intervention to the cluster level–require special consideration when running a power analysis. First, since multiple interventions are being tested, this means multiple, different anticipated effect sizes, which leads to different required sample sizes to detect these effects. Given this, it is advisable to run a separate power analysis for each intervention tested: one calculation for an individually randomised intervention, and one for a cluster randomised intervention.

Second, we need to address the cluster-level allocation of treatment of the whole-plot factor. For the reader already familiar with the chapters on cluster randomised trials and sample size and power, you may have already guessed that Intra-Cluster Correlation (ICC) is a relevant quantity for split plot trials. Since the hard-to-change, whole-plot factor is allocated at a cluster level, the units within each cluster may share characteristics that cause some correlation in their outcomes. In order to address this correlation, the sample size for the whole-plot factor needs to be proportionally adjusted using the design effect, which takes into account the correlation among units (the ICC) and the number of units per cluster.

Analytical approach

In many ways, split plot trial evaluations follow the same steps as a more traditional randomised trial: (1) data for outcomes and covarying variables are collected at two intervals (at least): baseline and endline; (2) treatment allocation is determined through a transparent and documented randomisation protocol; (3) the trial is closely monitored for attrition, treatment fidelity, spillover, and possible confounding factors; and (4) results are analysed, comparing outcomes between treatment and control groups. However, split plot trials differ from completely randomised trials in two key ways: the two-part (or more) nested randomisation protocol, and the associated multiple levels to account for in the analysis and interpretation of results. As this section will explain, running a simple OLS regression with two treatment inputs is inadequate: first, we need to account for interaction effects, and

second, we need to account for two levels of variance–at the whole-plot level and split-plot level–which has implications for our standard errors and significance reporting.

Below is a notational model of a 2x2 split plot trial, describing the predicted effects on output y, with whole-plot treatment i and split-plot treatment j. The treatment variables can take the values of 1 or 0. with 1 meaning treatment is received:

$$y_{ijk} = a + \beta_1 x_i + \beta_2 x_j + (\beta_1 \cdot \beta_2) x_{ij} + \delta_{k(i)} + \varepsilon_{ijk}$$

Let's break this down:

- a – The constant, or the predicted outcome under control (receiving neither treatment)
- $\beta_1 x_i$ – Estimated effect of when whole-plot treatment i = 1, and split-plot j= 0
- $\beta_2 x_j$ — Estimated effect of when whole-plot treatment i= 0, and split-plot j= 1
- $(\beta_1 \cdot \beta_2)\, x_{ij}$ – Estimated additional effect of when whole-plot treatment i = 1 and split-plot treatment j= 1, also known as the interaction effect
- $\delta_{k(i)}$ – Error term of whole-plot, taking into account the impact of split-plot treatment allocation
- ε_{ijk} – Error term for the split-plot, or the individual error term.

Let's turn to the interaction effect, then we will briefly address the two error terms.

Challenge one: Accounting for the interaction effect

An interaction means that the presence of both treatments can create an additional effect, beyond the sum of the individual treatments (for a common example of interaction effects, consider how certain medications can interact with other drugs like alcohol to increase or dampen effects).

Analytically, interactions are actually not much of a challenge to address. As implied in the notational model above, an interaction term can be understood as multiplying two variables together; in our case, that is our two treatments. Since our treatments can only take values of 1 and 0, the interaction effect is straightforward to understand:

[A= 1, B = 1] → 1*1=1

[A= 0, B = 1] → 0*1=0

[A= 1, B = 0] → 1*0=0

[A= 0, B = 0] → 0*0=0

Our simple 2x2 model produces four possible treatment combinations, and of those four, only one (highlighted in orange) could produce an interaction effect, where (A*B) =1.

A common deployment of interactions is in subgroup analysis. For example, many studies anticipate that an intervention will have differing effects for women compared with men. In this case, the team would interact the treatment with gender: treatment_a*gender. Below is an example table output of an invented treatment A on outcome Y, with subgroup analysis by gender (female = 1):

	Dependent variable:
	Outcome Y
treatment_a	12.80
	(0.49)
gender	-6.32***
	(1.06)
treatment a*gender	-7.66***
	(1.08)
Constant	39.96***
	(5.92)
Observations	241
Note	*p**p***p<0.01

Above, we can see that the effect of the treatment is attenuated among women (the highlighted -7.66 estimate). So while the treatment has a positive effect for participants overall (12.80 predicted effect for men), we can expect to see a more modest effect for women (12.80 – 7.66 = 5.14 predicted effect for women).

Let's return to our example of improving college attendance through a two-part intervention. Recall that the whole-plot intervention is whole-class incentives and the split-plot is text messages to parents. A model output for this might look like the following:

	Dependent variable:
	attendance
incentives	6.11*
	(0.73)
sms	4.25***
	(0.22)
incentives*sms	2.80**
	(1.28)
Constant	143.99***
	(22.4)
Observations	220
Note	*p**p***p<0.01

A quick read of the above estimates shows that both interventions–incentives and text messages–appear to have a positive, statistically significant effect on attendance compared with the control. The interaction effect–the additional effect when a student is exposed to both interventions–is relatively modest, but still statistically significant at the p-value level of 0.05.

Challenge two: Managing two levels of variance

As shown in our notational model above, our 2x2 split plot model has two error terms: one associated with the higher order, whole-plot factor, and the other with lower order split-plot. These error terms represent all the normal things you would expect–measurement error, variability from uncontrolled factors, variability in the experimental units to which the treatments are applied to, general background noise–but crucially, these error terms correspond to specific levels in the experiment. One of the key differences between these two error terms is that the split-plot error has more degrees of freedom, owing to the fact that the split-plot error term by definition incorporates more observations than the whole-plot error term. Since we have more independent observations for the split-plot factor (and thus more information), our error term for this level is likely smaller than the error term for the whole-plot.

If we ignore the nesting of treatments and analyse the model as if both treatments were completely randomised (as in a traditional factorial model), our statistical analysis software will assume a single error term, and calculate standard errors for each estimated effect based on residuals (or deviations) from a single sample mean. This does not reflect what we already know–that there are two error terms corresponding to each treatment level. This combined error term will likely sit somewhere between a larger error term (associated with the whole-plot) and the smaller error term (split-plot). The problem with this is we are leaving one of our core benefits of split plot design on the table: a more precise effect estimate for the split-plot factor. We are also potentially underreporting the standard error for the whole-plot factor, possibly leading us to Type I error for the whole-plot factor and Type II error for the split plot factor.

Estimating effects

Common approaches to modelling split plot trials are ANOVA or linear mixed effects modelling, where terms for the whole-plot, sub-plot, and their interactions are included in the model (Zhao et al, 2018). Note however that the confidence intervals for the whole-plot and sub-plot terms will need to be calculated separately, as each will have their own, different statistical error terms.

Conclusion

Split plot trials offer a practical design to test multiple interventions in combination with one another, with the flexibility to allow randomisation restrictions for those hard-to-change factors. When analysed carefully, split plot trials can also produce more precise effect estimates for the split-plot factor, potentially increasing our trial's efficiency. Given their design of incorporating multiple levels of randomising treatment, split plot trials require special consideration across the randomisation stage, sample size calculations/power analyses stage, and the modelling and analysis stage. The following table summarises these special considerations:

Table 1 Considerations for split plot trials

Trial Stage	Considerations for Split Plot Trials
Randomisation	Randomisation takes place in two (or more) steps: 1. The hard-to-change factor is randomly allocated to clusters (whole-plot). 2. Then the easier-to-change factor is allocated to units within those clusters (split-plot).
Sample size/ Power analysis	Each intervention should have a separate power analysis run to determine the needed sample size, taking into account each intervention's anticipated effect size. For the whole-plot intervention allocated to clusters, the power analysis should factor in Intra Cluster Correlation (ICC) and its associated design effect.
Modelling and analysis	To detect combined effects, intervention variables should be interacted with one another as part of the modelling specification. Modelling needs to account for two (or more) error terms: one at the whole-plot level, and another at the split-plot level. Not accounting for this will likely lead to an artificially high standard error for the split-plot term and an artificially low standard error for the whole-plot term.

References

Altman, N., & Krzywinski, M. (2015). Split plot design. Nature Methods, 12(3), Article 3. https://doi.org/10.1038/nmeth.3293

Burgess, S., Metcalfe, R., & Sadoff, S. (2021). Understanding the response to financial and non-financial incentives in education: Field experimental evidence using high-stakes assessments. Economics of Education Review: 85. https://doi.org/10.1016/j.econedurev.2021.102195.

Chande, R., Luca, M., Sanders, M., Soon, X.-Z., Borcan, O., Barak-Corren, N., Linos, E., Kirkman, E., & Robinson, S. (2017). Increasing attendance and attainment among adult students in the UK: Evidence from a field experiment. Behavioural Insights Research Centre for Adult Skills and Knowledge. https://www.bi.team/wp-content/uploads/2019/01/Texting-students-ALERT-working-paper-2017.pdf

Fryer, Jr., R.G., Levitt, S.D., List, J., and Sadoff, S. (2022). Enhancing the Efficacy of Teacher Incentives through Framing: A Field Experiment. American Economic Journal: Economic Policy, 14 (4): 269-99.

Hume, S., O'Reilly, F., Groot, B., Chande, R., Sanders, M., Hollingsworth, A., ... & Soon, X. (2018). Improving engagement and attainment in maths and English courses: Insights from behavioural research. Retrieved August, 8, 2018.

Luis A. Cortes, James R. Simpson, & Peter A. Parker. (2018). Response Surface Split Plot Designs: A Literature Review. Quality and Reliability Engineering International, 34(7). https://doi.org/oai:casi.ntrs.nasa.gov:20190029242

Parker, P.A., Kowalski, S.M. and Vining, G.G. (2006). Classes of Split-plot Response Surface Designs for Equivalent Estimation. Quality and Reliability Engineering International: 22, 291-305. https://doi.org/10.1002/qre.771

Yates, F. (1933). The principles of orthogonality and confounding in replicated experiments. (With Seven Text-figures.). The Journal of Agricultural Science, 23(1), 108-145. doi:10.1017/S0021859600052916

Zhao, A.,Ding, P., Mukerjee, R., and Dasgupta, T. (2018). Randomization-based causal inference from split-plot designs. The Annals of Statistics: 46(5), 1876-1903.

Chapter 14. Multi-arm trials and mega studies

The simplest form of randomised trial has two arms; one treatment and one control. As we've said elsewhere, the control group often does not receive no intervention at all; rather, they continue with business as usual so that we can have a fair test of whether or not the new intervention represents an improvement on what participants would otherwise receive.

Often, however, we will have multiple interventions that could be tested to achieve a particular outcome, and which target the same cohort of participants and have the same (or very similar) eligibility criteria.

Where this is the case, it can be more efficient to conduct a multi-arm trial, in which two or more treatments are tested against the control. There are arguments both for and against multi-arm trials that warrant consideration.

Comparability

One of the major benefits of a multi-arm trial is that it increases comparability between the effectiveness of the interventions tested, compared with running two (or more) independent two armed studies.

This comparability comes from the fact that the two interventions are being tested on the same sample, with the same characteristics, in the same contexts, and at the same point in time. Comparing the interventions in this context means that whichever performs the best in the trial represents your 'best bet' with some level of confidence.

This comparability point is not perfect, however, and depends a lot on the extent to which interventions are truly comparable. For example, if two interventions would ordinarily have different eligibility criteria, a multi-arm trial will either involve expanding the eligibility of one, restricting the eligibility for another of the interventions. This needn't be problematic – you will still be capturing an accurate measure of the effect of the treatments on those people who are included in the trial, but that may limit the generalisability of your findings. Another source of a lack of comparability might be the intensity or duration of the intervention. Let's consider two programmes of support for people who have previously experienced rough sleeping that aims to get them back into work. If one programme costs half as much, and/or takes half as long, then the trial will need to be longer than necessary for that intervention in order to capture the effects of the intervention. Similarly, it is difficult to make comparisons of value for money based on the effect sizes of the trial along where there are very different costs of the two interventions.

Logistical

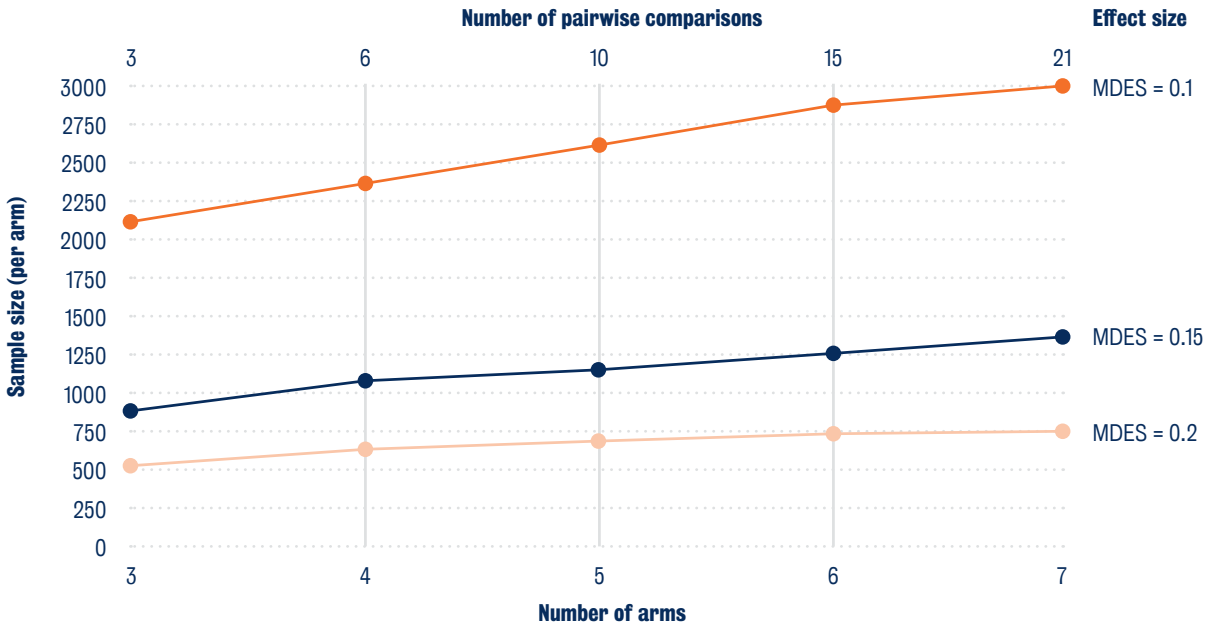
The logistical advantages of running a multi-armed trial are that fewer units of randomisation need to be recruited, compared to running two trials that only test one intervention, because only one control group is needed, rather than two. This reduces the effort of recruitment, but also, relative to running two trials, reduces the costs of data collection, and can reduce the costs associated with analysis, and getting ethical approval.

However, as with comparability, there are downsides to a multi-arm trial too. In particular, individual participants, or institutions (in the case of a cluster randomised trial), must be recruited, not knowing whether they will receive intervention 1, intervention 2, or neither. This is a different prospect to a standard trial, because while they will often have a higher probability of receiving an intervention, they may have a preference for which intervention they receive, and perhaps not be interested in receiving the other.

Statistical power and Multiple comparisons

As we have seen in Chapter 6 when thinking about statistical power, it is efficient to have one control group and test multiple intervention groups against it. However, there is also a trade-off because adding intervention arms also increases the number of comparisons that are being made, particularly if you are formally comparing the efficacy of one intervention to that of another, and not simply comparing the efficacy of both interventions separately compared to the control. This necessitates (or makes more likely, depending on the guidance you follow), making adjustments for multiple comparisons – meaning that the sample will need to increase to achieve the same level of statistical power for a given minimum detectable effect size. The rate at which this trade-off exists can be seen in the diagram below.

Figure 1 Sample size by number of intervention arms



The graph shows how increasing the number of arms of the multi-arm trial results in an –almost linear- increase of the per-arm sample size for a given effect size following the Bonferroni correction. As the number or arms increases, the number of pairwise comparisons increases multiplicatively by a factor of: (arms*(arms-1))/2. For example, assuming we need to power a trial to detect differences between 4 arms: A, B, C and D (a control and 3 interventions), the number of pairwise comparisons or hypothesis tests will rise to 6 (A vs. B; A vs. C; A vs. D; B vs. C; B vs. D; C vs. D). If instead we only cared about comparisons between the control and each treatment arm, we would assume the number of pairwise comparisons as 3 in the above scenario and adjust our sample size accordingly.

The trade-off between sample size and number of arms under Bonferroni results in an increase of the required sample size within each MDES scenario; in contrast, as evident from the graph, the change in sample size between MDES scenarios is asymptotic, which is a characteristic of the power calculation's exponential nature. It is important to note that if we equally care about the expected difference between treatment arms, and that difference is expected to be less than their difference to the control, the sample size will require a further upward adjustment as a result of the effective change in the MDES of interest.

As explained in earlier chapters, the BH adjustment has no closed form which makes it analytically difficult to use when powering a trial. Therefore, although power calculations should involve the Bonferroni correction for ease, other things equal, the implied post-hoc power will be higher with respect to a BH adjustment.

Example of a multi-armed study

An example of a multi-arm study is the FSQ1 trial that we are carrying out as part of a consortium for the Department for Education as part of their Multiply Programme of trials. This trial was originally intended to be two trials, each testing a different intervention. The first trial focused on testing a mathematics mastery intervention, which (in common with other mastery approaches), encourages learners to master one topic before moving on to the next. The second was to test a contextualised curriculum, in which mathematical calculations were related to features of people's everyday lives, to make them more relevant and easier to follow. Both interventions were intended to be carried out with students learning for Foundation Stage Qualifications Level 1 in Numeracy and delivered in adult learning providers such as Further Education Colleges. Given the small number of FE colleges (In total around 280), and the need for large samples because of cluster randomisation, it was decided to merge the two trials together. Although the interventions differed slightly in terms of their duration and in their potential outcomes per the theory of change, the core outcome of numeracy was the same, and the reduced total sample size requirement for the multi-armed trial meant that merging the two trials was deemed worthwhile.

Mega studies

A mega study is a special case of the multi-arm trial, in which large numbers of interventions are simultaneously tested against a control condition. This approach has been popularised by the Behavioural Scientists Katherine Milkman and Angela Duckworth, as part of their Behavior Change for Good initiative.

In the Behavior Change for Good Initiative, an organisation with a challenge (for example a Gym Company or public health organisations aiming to increase vaccine intake), identifies a common outcome measure (number of trips to the gym in six months; vaccine uptake), and conditions for the intervention – for example that it must be delivered by text message, and must cost less than $2 per participant, and these conditions are disseminated to a large number of behavioural scientists, who form small teams and submit their chosen intervention. Then, large numbers of participants are recruited (typically in the tens of thousands), and randomised to receive one of the interventions, or the control condition. This creates a study with potentially dozens of different conditions all being tested at the same time.

Mega studies have many of the same challenges as multi-arm studies, writ large – notably that they require large sample sizes, and have a coordination problem – different interventions need all to be ready to go at the same time. Issues of multiple comparisons also become more acute with each additional intervention arm.

It is for this reason that this methodology has thus far been mainly used to test behavioural interventions like nudges which are light touch, low cost, and amenable to standardisation. However, there is no reason in principle that they could not be tested for more substantial interventions, like training (in which large numbers of workers could be randomised to receive different trainings), or interventions that are already delivered to large numbers of people.

Example of a mega-study

Mega-studies are relatively new as an approach to testing interventions. One megastudy, led by Katherine Milkman from the University of Pennsylvania's Wharton School, focused on how to increase uptake of flu vaccines. The study, which had 47,306 participants, tested 19 different text message based interventions simultaneously. The interventions on average increased response rates by 5 per cent. The best-performing intervention in our study reminded patients twice to get their flu shot at their upcoming doctor's appointment and indicated it was reserved for them. This intervention increased response rates by 11 per cent. The approach allowed for large numbers of interventions to be tested quickly in a single time period, and for insights to be produced that could be deployed to increase uptake of Covid-19 vaccines. Several interventions significantly outperformed the control, suggesting that if attenuation of effects of the main due to repetition was identified when the intervention was scaled, these other interventions might also be worth deploying. [1]

Conclusion

Multi-arm trials–popularized by large-scale behavioural science mega-studies–can be an effective design for evaluating multiple interventions within a single study. By utilizing a shared control group, these trials enable direct comparisons of effect sizes amongst different intervention carried out in the same context. This approach can be cost-effective, as it minimizes the need for multiple independent trials and streamlines data collection and analysis. However, multi-arm trials present several statistical and operational challenges, including the need for larger sample sizes to maintain power and the complex management of coordinating the simultaneous rollout of multiple interventions. References

References

An Adapted Mastery Approach and Contextualised Curriculum for Functional Skills Qualification Level 1 Trial Specification (2024). Department for Education.
https://multiply.etioglobal.org/hubfs/Multiply/Multiply%20Trials%20Files/FSQ%20L1%20Maths%20Combined%20Trial%20Specification.pdf

Milkman, K. L., Patel, M. S., Gandhi, L., Graci, H. N., Gromet, D. M., Ho, H., Kay, J. S., Lee, T. W., Akinola, M., Beshears, J., Bogard, J. E., Buttenheim, A., Chabris, C. F., Chapman, G. B., Choi, J. J., Dai, H., Fox, C. R., Goren, A., Hilchey, M. D., … Duckworth, A. L. (2021). A megastudy of text-based nudges encouraging patients to get vaccinated at an upcoming doctor's appointment. Proceedings of the National Academy of Sciences, 118(20), e2101165118.
https://doi.org/10.1073/pnas.2101165118

Chapter 15. Winner stays on trials

In chapter 14, we considered multi-arm trials when we want to test multiple interventions in a single trial. However, sometimes the sample available for any given experiment may not be large enough to allow for testing multiple interventions, but the circumstances of the trial recur, allowing for multiple similar trials to be conducted sequentially.

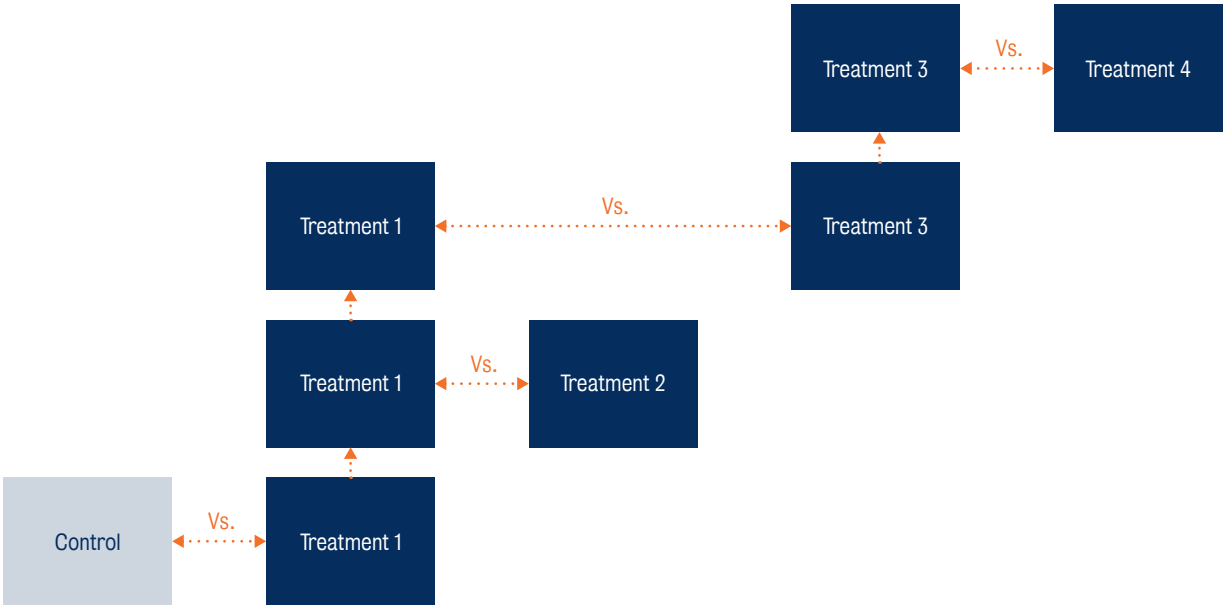
This is a common phenomenon with communications trials, but also occurs in any context where there are new cohorts of potential participants arriving at regular intervals – as is the case with probation services; in residential care settings for younger or older people; in schools (with less frequency), and so on.

In this context, we might want to test what is the most effective method of supporting these people, but not to wish to keep running experiments in which new interventions are compared against the control; and the randomisation to several conditions within any one cohort is likely to be unsuccessful because of the small sample.

Where this is the case, we can conduct a multi-stage, winner-stays-on research design, as shown in the figure below. In this kind of design, the first cohort is subject to a two-armed randomised trial, in which the previous business as usual is tested against a new intervention. In the context of communications trials, this means that the usual messages are sent to half of the participants, and a new message, perhaps making use of a behavioural insight, is sent to the other half.

After this round of communications is conducted, the efficacy of the two interventions is compared to each-other, and analysis is conducted. After this test is concluded, we determine which intervention 'won' that round of the test, and then proceed to the second, in which the winner is tested against treatment 2, in the second cohort of communications. If treatment 1 triumphs over treatment 2, then the third trial takes places in which treatment 3 is tested against treatment 1, and so on.

Figure 1 Winner stays on



This approach to trials is, of course, limited in its uses – to situations where there are repeated cohorts, often quite quickly after each other. However, for interventions that are focused on communications, or where these cohorts do occur rapidly, they do allow you to learn quickly and to discard interventions that are not working, without the logistical complexity of running a multi-armed trial that might not be appropriate.

They are also ethical benefits to the winner-stays on trial. For example, if after the first cohort, we know that treatment 1 is substantially outperforming the control, we can remove the control from consideration before we go into cohort two. This means, that we are able to maintain the sense of equipoise – of not knowing which intervention is best performing – if the first cohort has yielded a sufficiently decisive result.

UKRI Metascience Peer Reviewer Completion

An example of a multi-stage 'winner stays on' design that we are currently running in collaboration with the UKRI Metascience Unit. This experiment is testing nudging messages sent with the intent of increasing peer reviewer report completion rate of individuals invited by UK Research Councils to review applications for research funding. In this design, nudge messages are assigned into four groups: obligation nudge, pro-social nudge, self-interest nudge, and a control message. Each message is tagged with its assigned nudge type, and all proposals sent for review include an indicator of the assigned nudge. In the first stage, two conditions–control and treatment 1 (eg, obligation nudge)–are tested against each other. The intervention with the better performance, based on a defined outcome of percentage of reviewers signing up, is deemed the 'winner.' In the subsequent stage, the winning condition is tested against treatment 2 (eg, pro-social nudge), and this process continues until a final 'winner' of the tournament is identified. This iterative process allows for a systematic evaluation of interventions to identify the most effective nudge, while complementing the rolling nature of the contact and signing-up process.

Conclusion

The Winner Stays On trial design can be deployed to test multiple interventions sequentially when cohorts of participants are regularly introduced over time. This is particularly useful for trials with smaller sample sizes that are not suited for a multi-armed study, but where we are still interested in testing multiple interventions for the same outcome. While its applicability is limited to contexts with frequently recurring cohorts, this design provides valuable ethical and practical advantages by allowing researchers to eliminate underperforming interventions early and streamline trial logistics.

Annex 1. Trial Protocol Template

1. Trial protocol

[Title of project]

Version	Date	Reason for revision
1.0 [original]		[Original – as registered on OSF]

Principal Investigator:

Contact Email:

Open Science Framework Reference Number

2. Project planning

Background and Problem Statement

Key Personnel

Table 1 Key personnel

Name	Organisation	Role in project

Timetable

Table 2 below provides an overview of the timelines for the evaluation.

Table 2 Mapping of timeline for evaluation activities

Task	Timing

3. Intervention and theory of change

Table 3 shows the TIDieR framework for the evaluation.

Table 3 TIDieR framework

Name	
Why	
Who (recipients)	
What (materials)	
What (procedures)	
Who (provider)	
How (format)	
Where (location)	
When and how much (dosage)	
Tailoring	
Control condition	

Details of the intervention as-implemented and experienced by participants, as well as business-as-usual support accessed by both treatment and control participants, will be explored via the process evaluation and documented in the report of the trial.

4. Impact evaluation

Research Questions

Design

Outcome measures

Table 4, below, gives the outcomes and associated outcome measures for the impact evaluation.

Table 4 Outcome measures

Primary outcome	Variable
	Measure(s) (instrument, scale)
Secondary outcome	Variable
	Measure(s) (instrument, scale)
Secondary outcome	Variable
	Measure(s) (instrument, scale)
Secondary outcome	Variable
	Measure(s) (instrument, scale)
Secondary outcome	Variable
	Measure(s) (instrument, scale)

Please note that the research questions regarding access to public services and the justice system will be addressed in the economic evaluation (further details available in section 12).

Sample size calculations

Table 5 Minimum Detectable Effect Size Calculations

Unit of randomisation	
Alpha	
Power	
Baseline-Endline Correlation (Housing Stability)	
Total sample size across both arms	
MDES (Cohen' d)	

Participants

Inclusion and exclusion criteria

Randomisation

Data collection

Table 6 Data collection

Data item	When	Data collector	Data protection

Strategies for participant engagement and retention

- Analytical Strategy
- Descriptive Statistics
- Primary Analysis
- Secondary Analysis
- Missing Data
- Multiple Comparisons
- Exploratory Analysis

6. Implementation and Process Evaluation

Aims

Research Questions

Data Source and Data Collection Tools

Qualitative Methods

Table 7 IPE data collection

Sample	Method	Delivery	Time

Semi-structured interviews with applicants

Quantitative Methods

Data analysis and synthesis

7. Economic Evaluation

8. Ethics

9. Registration

This protocol will be registered with the Open Science Framework once ethical clearance is obtained and prior to data collection commencing. The protocol will be updated with the reference link once registration is complete.

10. Risks

Table 1 Risk log

Risk/issue	Mitigation

11. Data Protection

Annex 2

Sample size calculations for multiple comparisons

Table 1 Sample size calculations for multiple comparisons

Effect Size	Baseline-Endline Correlation	No. of Conditions	Multiple Comparisons*	Required Sample Size (Per arm)
0.1	0	1	No adjustment	1570
0.1	0.3	1	No adjustment	1429
0.1	0.5	1	No adjustment	1178
0.2	0	1	No adjustment	393
0.2	0.3	1	No adjustment	358
0.2	0.5	1	No adjustment	295
0.3	0	1	No adjustment	175
0.3	0.3	1	No adjustment	159
0.3	0.5	1	No adjustment	131
0.4	0	1	No adjustment	99
0.4	0.3	1	No adjustment	90
0.4	0.5	1	No adjustment	74
0.1	0	2	Bonferroni	1902
0.1	0.3	2	Bonferroni	1730
0.1	0.5	2	Bonferroni	1426
0.2	0	2	Bonferroni	476
0.2	0.3	2	Bonferroni	433
0.2	0.5	2	Bonferroni	357
0.3	0	2	Bonferroni	212
0.3	0.3	2	Bonferroni	193
0.3	0.5	2	Bonferroni	159
0.4	0	2	Bonferroni	119
0.4	0.3	2	Bonferroni	109
0.4	0.5	2	Bonferroni	90
0.1	0	2	Benjamini-Hochberg	1641
0.1	0.3	2	Benjamini-Hochberg	1493
0.1	0.5	2	Benjamini-Hochberg	1231
0.2	0	2	Benjamini-Hochberg	410
0.2	0.3	2	Benjamini-Hochberg	373
0.2	0.5	2	Benjamini-Hochberg	308
0.3	0	2	Benjamini-Hochberg	183
0.3	0.3	2	Benjamini-Hochberg	167
0.3	0.5	2	Benjamini-Hochberg	137
0.4	0	2	Benjamini-Hochberg	102
0.4	0.3	2	Benjamini-Hochberg	93
0.4	0.5	2	Benjamini-Hochberg	77
0.1	0	3	Bonferroni	2094
0.1	0.3	3	Bonferroni	1906

0.1	0.5	3	Bonferroni	1571
0.2	0	3	Bonferroni	524
0.2	0.3	3	Bonferroni	477
0.2	0.5	3	Bonferroni	393
0.3	0	3	Bonferroni	233
0.3	0.3	3	Bonferroni	212
0.3	0.5	3	Bonferroni	175
0.4	0	3	Bonferroni	131
0.4	0.3	3	Bonferroni	120
0.4	0.5	3	Bonferroni	99
0.1	0	3	Benjamini-Hochberg	1649
0.1	0.3	3	Benjamini-Hochberg	1501
0.1	0.5	3	Benjamini-Hochberg	1237
0.2	0	3	Benjamini-Hochberg	415
0.2	0.3	3	Benjamini-Hochberg	378
0.2	0.5	3	Benjamini-Hochberg	311
0.3	0	3	Benjamini-Hochberg	185
0.3	0.3	3	Benjamini-Hochberg	168
0.3	0.5	3	Benjamini-Hochberg	139
0.4	0	3	Benjamini-Hochberg	104
0.4	0.3	3	Benjamini-Hochberg	95
0.4	0.5	3	Benjamini-Hochberg	78
0.1	0	4	Bonferroni	2231
0.1	0.3	4	Bonferroni	2030
0.1	0.5	4	Bonferroni	1673
0.2	0	4	Bonferroni	558
0.2	0.3	4	Bonferroni	508
0.2	0.5	4	Bonferroni	419
0.3	0	4	Bonferroni	248
0.3	0.3	4	Bonferroni	226
0.3	0.5	4	Bonferroni	186
0.4	0	4	Bonferroni	140
0.4	0.3	4	Bonferroni	127
0.4	0.5	4	Bonferroni	105
0.1	0	4	Benjamini-Hochberg	1660
0.1	0.3	4	Benjamini-Hochberg	1511
0.1	0.5	4	Benjamini-Hochberg	1245
0.2	0	4	Benjamini-Hochberg	417
0.2	0.3	4	Benjamini-Hochberg	379
0.2	0.5	4	Benjamini-Hochberg	313
0.3	0	4	Benjamini-Hochberg	186
0.3	0.3	4	Benjamini-Hochberg	169
0.3	0.5	4	Benjamini-Hochberg	140
0.4	0	4	Benjamini-Hochberg	105
0.4	0.3	4	Benjamini-Hochberg	96
0.4	0.5	4	Benjamini-Hochberg	79
0.1	0	5	Bonferroni	2336
0.1	0.3	5	Bonferroni	2126
0.1	0.5	5	Bonferroni	1752

0.1	0.5	5	Bonferroni	1752
0.2	0	5	Bonferroni	584
0.2	0.3	5	Bonferroni	532
0.2	0.5	5	Bonferroni	438
0.3	0	5	Bonferroni	260
0.3	0.3	5	Bonferroni	237
0.3	0.5	5	Bonferroni	195
0.4	0	5	Bonferroni	146
0.4	0.3	5	Bonferroni	133
0.4	0.5	5	Bonferroni	110
0.1	0	5	Benjamini-Hochberg	1668
0.1	0.3	5	Benjamini-Hochberg	1518
0.1	0.5	5	Benjamini-Hochberg	1251
0.2	0	5	Benjamini-Hochberg	418
0.2	0.3	5	Benjamini-Hochberg	380
0.2	0.5	5	Benjamini-Hochberg	314
0.3	0	5	Benjamini-Hochberg	187
0.3	0.3	5	Benjamini-Hochberg	170
0.3	0.5	5	Benjamini-Hochberg	140
0.4	0	5	Benjamini-Hochberg	106
0.4	0.3	5	Benjamini-Hochberg	96
0.4	0.5	5	Benjamini-Hochberg	80

The School for Government

The School for Government at King's College London is committed to creating a world governed with intelligence, integrity and innovation.

In an era defined by complex challenges and rapid change, we need a new approach to educating the current and future leaders who will guide our institutions through turbulent times. Our programmes are based on three core elements:

Partnerships

Our work is built on deep partnerships with government institutions in the UK and internationally. Through our connection with both the Policy Institute at King's and the wider university, we bring together practitioners, scholars, and students in a collaborative environment that bridges theoretical understanding with practical application. Learn government from those who have governed.

Practice

The school's foundation rests in the day-to-day reality of government and governance. Our teaching aligns with civil service principles while fostering a new breed of scholar – one who understands both academic rigor and practical implementation. This approach creates graduates who can navigate the complexities of modern governance while maintaining the highest standards of public service.

Performance

We emphasise excellence across all aspects of our programmes. Our diverse faculty includes both distinguished academics and experienced practitioners who bring real-world insights to the classroom. We've designed our administrative processes to be efficient and student-centred, ensuring that learning remains the primary focus of everyone who studies here.

The result is an educational experience that prepares people not just to understand government, but to transform it, with our graduates emerging ready to tackle the challenges of modern governance with both practical skills and sound judgement.

kcl.ac.uk/sfg