TASO Transforming Access and Student Outcomes in Higher Education

KING'S College LONDON

Report:

Evaluating complex interventions using randomised controlled trials[1]

Michael Sanders, King's College London

Julia Ellingwood, King's College London

Dr Eliza Kozman, TASO

August 2023

CONTENTS

Acknowledgements

This paper has been written by researchers at King's College London in partnership with colleagues from the Centre for Transforming Access and Student Outcomes in Higher Education (TASO). It is a version of a chapter of the Ochre Book, which is funded under the DUPLO project funded by the UK Cabinet Office.

1. INTRODUCTION

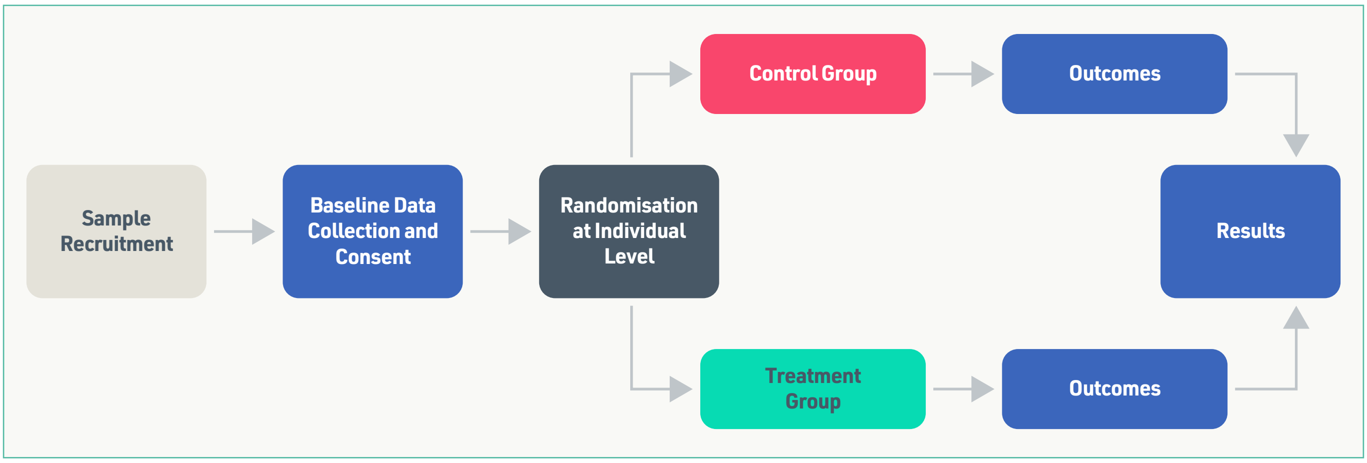
Randomised Controlled Trials (RCTs) are, in many ways, a straightforward tool for answering a straightforward question: "If I do X, what will happen to Y?". The question of "What Works" is answered through the stacking of lots of these questions together to ask "What X has the biggest effect on Y?".

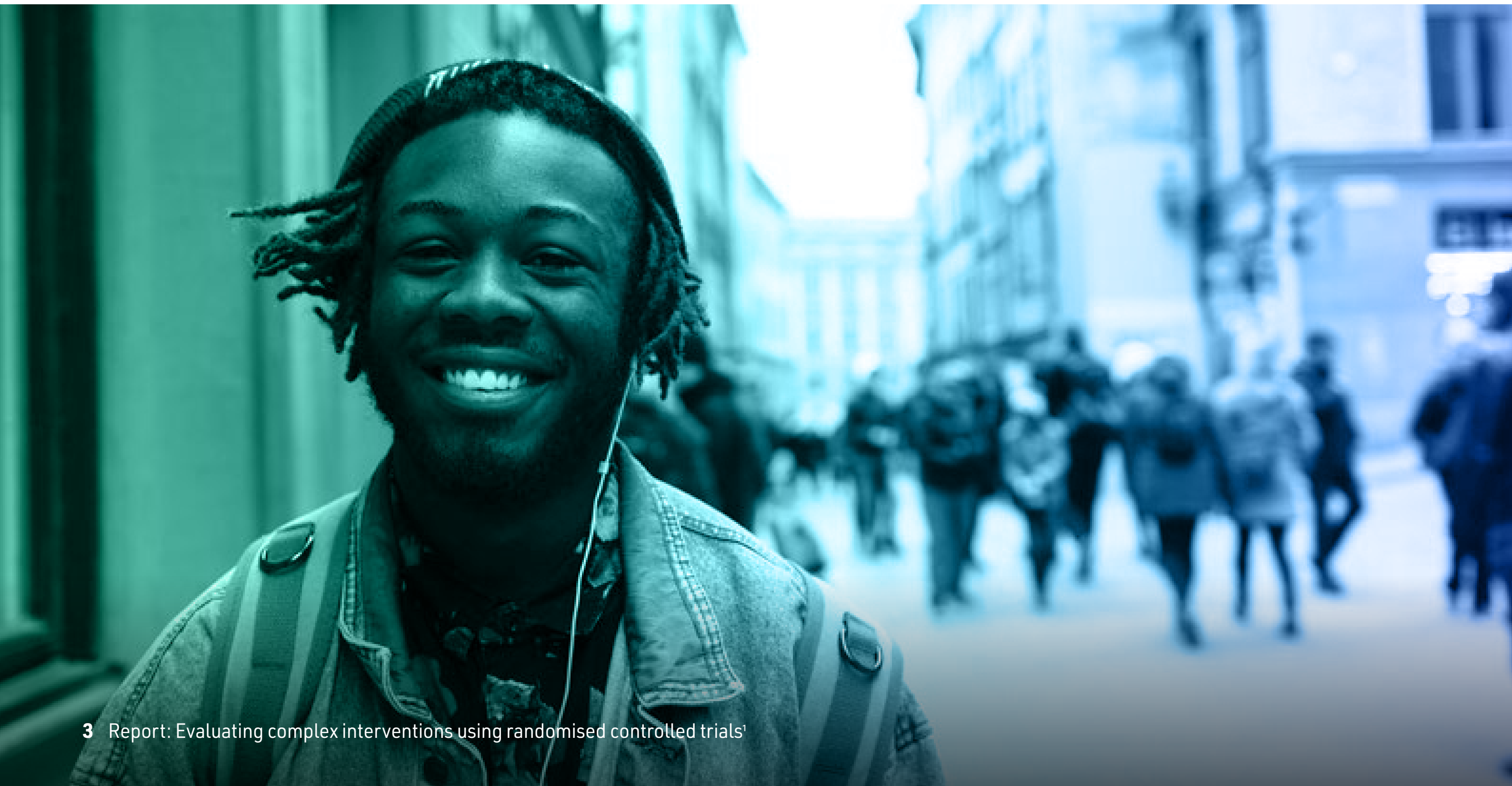
Although the process of running randomised trials can often be anything but simple, the idea, at least, is straightforward. The figure below shows the flow of a standard, parallel designed RCT.

There are many interventions that we might be interested in, which defy 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.


Sample Recruitment
Baseline Data Collection and Consent
Randomisation at Individual Level
Control Group
Outcomes
Treatment Group
Outcomes
Results



2. WHAT IS A COMPLEX INTERVENTION?

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:

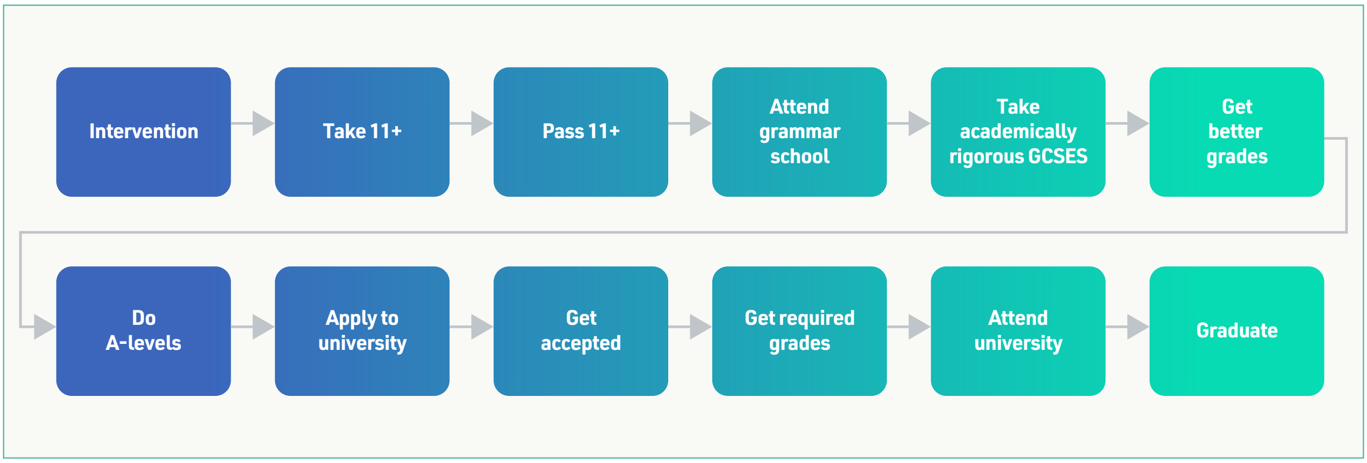
- Long causal chain interventions
- Multi-component interventions
- Multi-target interventions
- System-level interventions
- Evolving interventions

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).

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 also have 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.



Box 1. Testing the effect of bursary information on HE applications

TASO and King's College London collaborated to test whether or not 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 happen 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[5] 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.[6]

An example of an multicomponent intervention in higher education is given in Box 2.

[5] https://www.northyorks.gov.uk/no-wrong-door

[6] https://whatworks-csc.org.uk/evidence/evidence-store/intervention/intensive-family-preservation-services/

Box 2. 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 RCTs.

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.

Box 3. 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 & Mann, 2011; 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.

3. 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 enrollment. 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:

1. Elaborate a theory of change and specify the hypotheses to be tested.
2. Describe how emerging findings in the process evaluation of the trial will inform the refinement of the hypotheses.
3. 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 generalizability, 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 enrollment).

Box 4. 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 WP backgrounds in applying to highly selective universities. The K+ programme was evaluated in 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 (i.e., 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 enrollment 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.

Box 4. 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 COVID-19 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 tradeoffs

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.

Quasi-Experimental Designs: Matching and Difference in Differences

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 (e.g. 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.

DiD 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, i.e. 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.

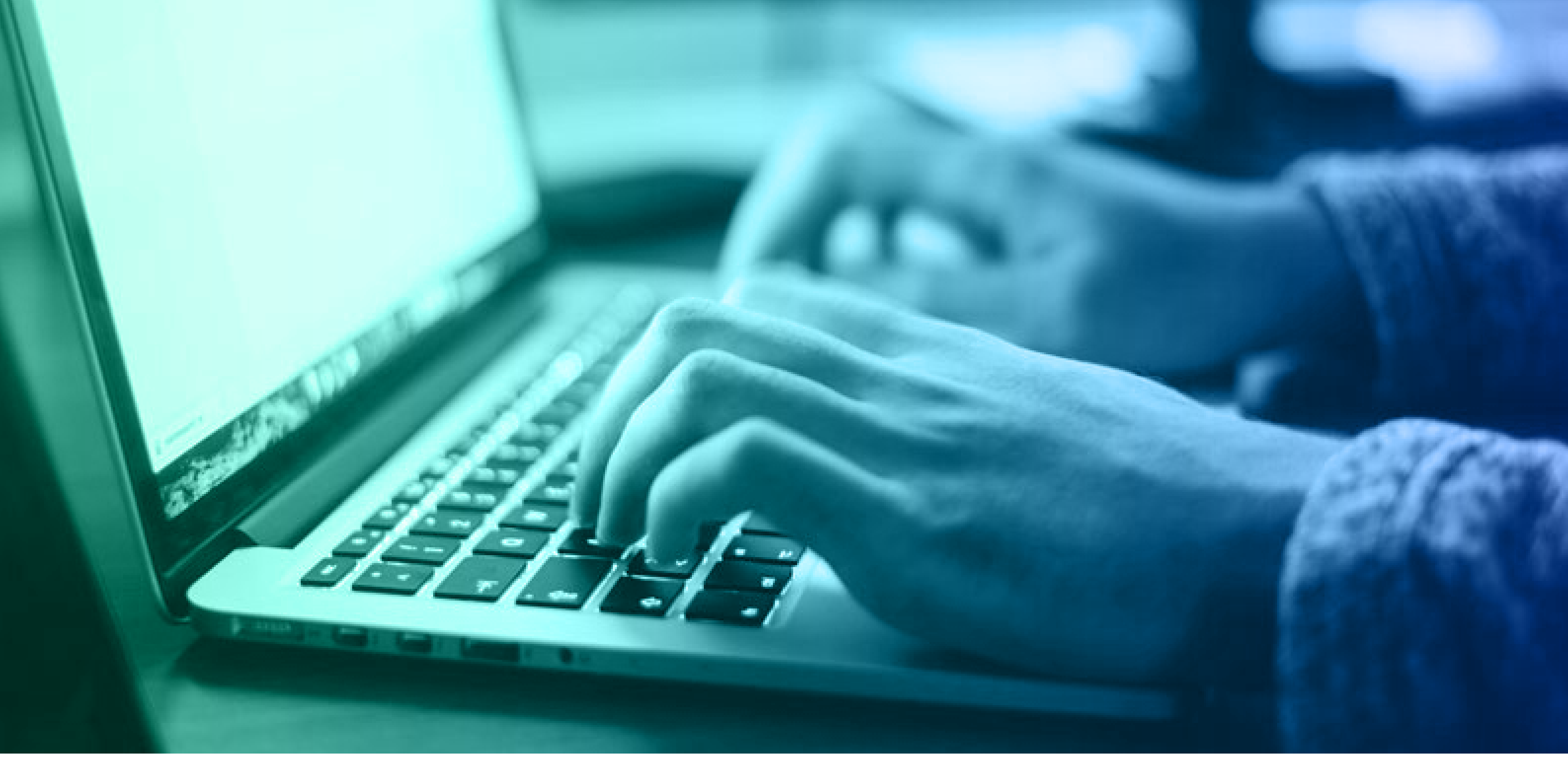
Multi-stage trial protocols

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 spuriously a statistically significant effect 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) (Kerr, 1998).

This rigidity is a core strength of randomised trials, and 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.

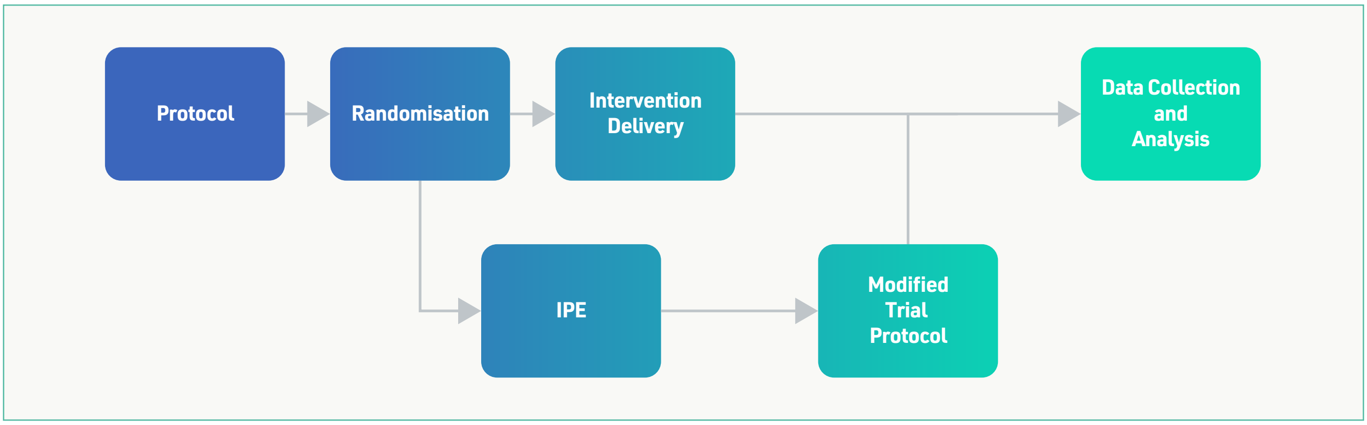
Over the course of the trial, various other forms of research, mostly in the form of implementation and process evaluation, can be conducted, which will give insights into how the intervention is conducted; which of several outcomes at the most likely to be effected, and which of many subgroups are most 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 prespecification, while allowing us to capture the complexity of the intervention and its effects through flexibility. The approach is shown in the diagram below.

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 most simple 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.

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.

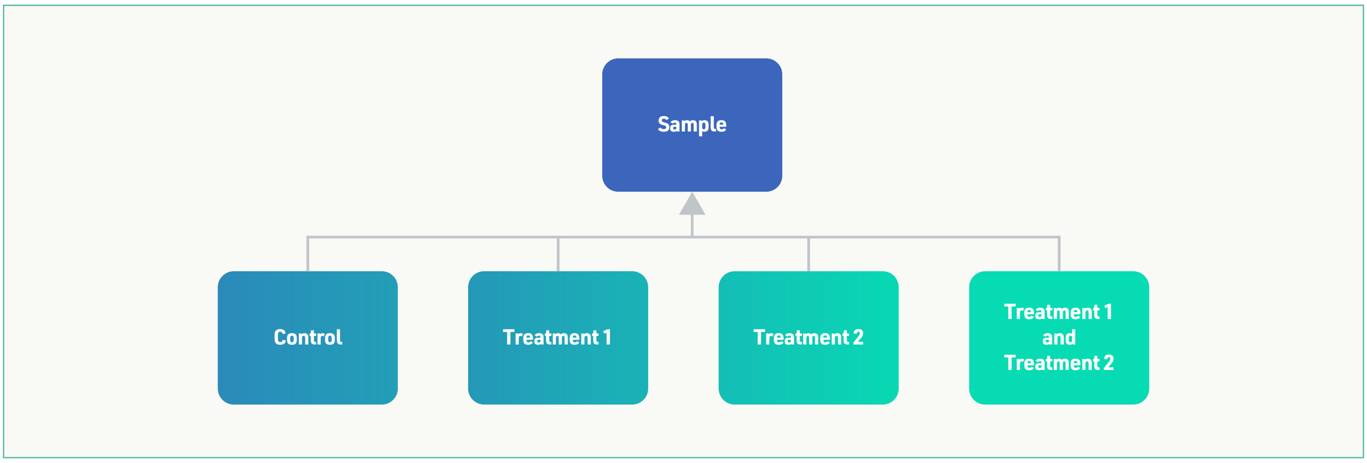
However, this approach is not without its problems. Typically we power our trials to detect effects of a reasonable size with 80% probability, relative to a control group that do not receive the intervention. In this case, this level of statistical power is designed to detect the difference between any 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% 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.

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

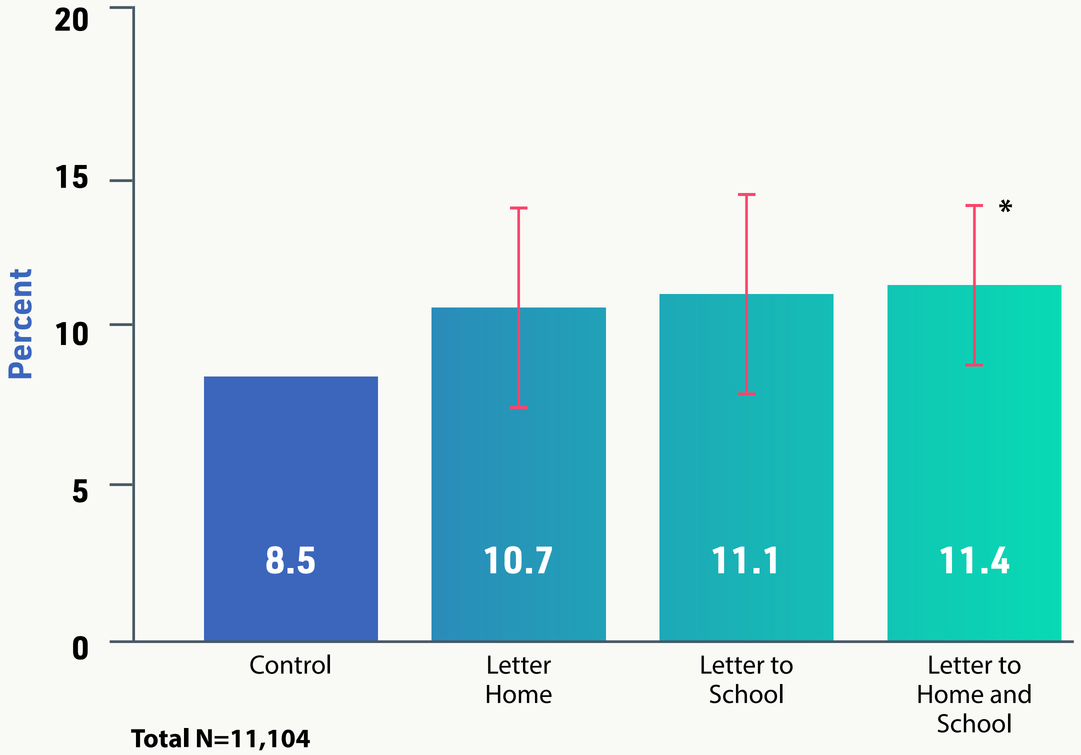


Powering for Interactions

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 four armed trial conducted by Sanders et al (2023), 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. The letters were written by relatable role models and the proportion of recipients applying to selective universities was measured. 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.

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.



Total N=11,104

*** p<0.001, ** p<0.01, * p<0.05, + p<0.1

Split Plot Trials

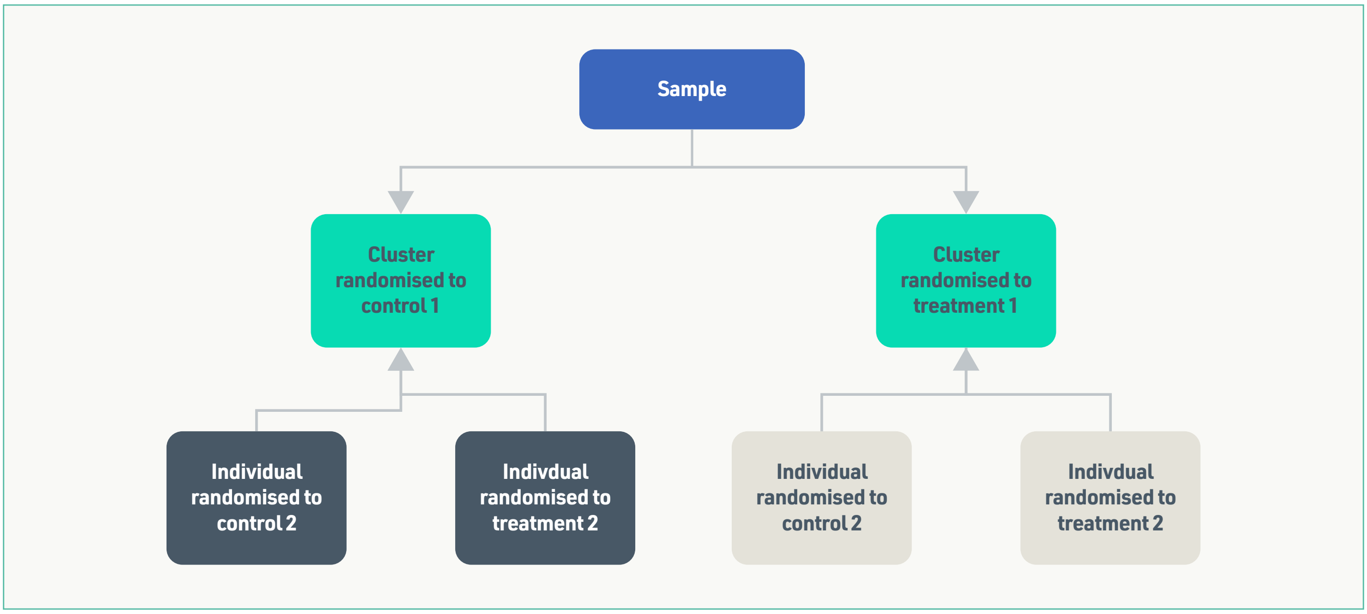
Split plot trials are a special case of factorial trials, where it is possible to test factorial designs with the same kind of grid structure as shown above, but in a way which is more statistically efficient than the examples given above.

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 relates to different levels of randomisation.

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.


Sample
Cluster randomised to control 1
Cluster randomised to treatment 1
Individual randomised to control 2
Indivdual randomised to treatment 2
Individual randomised to control 2
Indivdual randomised to treatment 2

Stepped wedge trials

Where we 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 might be a good option.

A stepped wedge trial is so called because the design of the trial resembles a staircase, in which participating units are randomised not to receive the intervention or not, as in a parallel trial, but instead to a time period at which they are to start the intervention.

In a stepped wedge, the intervention is gradually rolled out to all the units in the trial. This has the benefit of maintaining engagement from control units, and of allowing a scarce resource - such as time to deliver training, or support from consultants - to be rationed, without limiting the viability of the trial.

An example of a suite of stepped wedge trials is the evaluation of the Department for Education's Strengthening Families, Protecting Children programme[7], which sees three whole local authority models of change being rolled out to six new local authorities each over the course of several years. The local authorities who originated the interventions and practice models could not deliver the intervention to all six at once, as this is an intensive process. Instead, local authorities are supported to initiate the intervention in sequence, with a new local authority starting roughly every six months.

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 are therefore most likely to be appropriate when;

- The intervention can only be delivered to smaller proportion than 50% at any given point in time, usually quite a bit smaller.
- The intervention is anticipated to have short or medium term effects, so having some units receiving 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.

4. CONCLUSIONS

In this short paper 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.

[7] https://whatworks-csc.org.uk/research-project/family-valued-model-trial-evaluation/

5. 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.

Masset, E., Shrestha, S.,& Juden, M. (2021). Evaluating complex interventions in international development. *The Centre of Excellence for Development Impact and Learning (CEDIL).*

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.

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.*

TASO

Transforming Access
and Student Outcomes
in Higher Education

Evidence Quarter
Floor 4, Albany House
94-96 Petty France
London SW1H 9EA

info@taso.org.uk
taso.org.uk

TASO is an independent charity that aims to improve lives through evidence-based practice in higher education (HE). We support HE professionals through research, toolkits and evaluation guidance on what works best to eliminate equality gaps. We inform practitioners of the best available evidence and produce new evidence on the most effective approaches.
TASO is an affiliate 'What Works' centre and is part of the UK Government's What Works Movement.