Experimental Designs

What are experimental designs

When we rely on observational studies to evaluate the impact of a treatment or intervention, we are limited to analysing the variation that exists in the world, and much of that variation is tied to self-selection. Suppose we are interested in the effect of apprenticeships; youths opting for apprenticeship differ systematically (in ability, motivation, opportunity) from these who decline; how then can we detect the effect of apprenticeship itself? In experimental designs *we* specify the treatment, and, importantly, we *randomly* allocate who is to receive it – so eliminating selection effects. We can then compare 'treated' and 'untreated' participants in the expectation that outcome differences relate only to 'treatment', not selection into treatment. Participants can be individuals, groups, organisations, villages – indeed any unit of observation relevant to the question being posed.

The random allocation of participants is the one feature that marks experiments from other research designs. This entry focuses primarily on the role of such designs in providing a fair test of the effects of social interventions and focuses on what are known as Randomised Controlled Trials (RCTs). In doing so, we draw examples from intervention-oriented experiments designed to produce desired outcomes, *e.g.* to reduce delinquency. Social science experiments can of course be motivated without such a direct policy orientation (*e.g.*

experiments designed to identify determinants of helping behavior, obedience, or racial bias), but most of the discussion still applies to them.

We begin with the formal structure of generic experiments, but are alert throughout to consider the peculiar issues posed by the evaluation of *social* interventions.

Establishing causal relationships

Ascertaining cause and effect is difficult. In order to establish that A (the independent variable) causes B (the dependent variable), we need to be able to demonstrate i) that A comes before B, ii) that A covaries with B (for example, in the absence of A, B does not occur) and iii) B cannot be explained by something else. In the laboratory the researcher can manipulate interventions of interest in ways that enable the ruling out of most, if not all, competing explanations (see Wolbring, Sage Research Methods Foundation). An obvious disadvantage of the laboratory experiment is that the situation is – and is seen to be – artificial. A laboratory-analogue is the so-called field experiment.

Unlike laboratory experiments, field experiments take place where the behaviors of interest might naturally occur. The independent variable of interest is manipulated by the experimenter and randomisation is an important design component. For example, in 1965 Rosenthal and Jacobsen tested children in 18 classes within an elementary school. The test was a straightforward IQ test. However, in 'reporting' to the teachers the researchers said (falsely) that the test was designed to predict which children would 'bloom' academically, and they randomly assigned 20% of the pupils to their experimental condition, identifying them to their class teachers as pupils who would show 'unusual intellectual gains during the academic year'. Eight months later the children were retested; the researchers found that

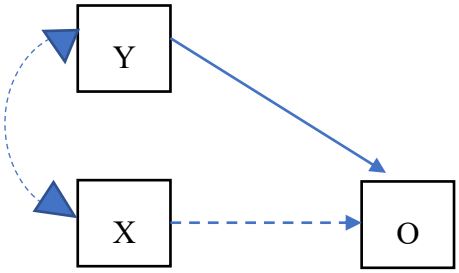
those in the experimental group showed significantly greater IQ gains than those in the control. The impact of teacher expectation on pupil performance was most marked among younger children, and there has been a rich seam of academic papers exploring the mechanisms that might account for Rosenthal and Jacobsen's findings. (Notice in passing that this 1965 research would have been unlikely to receive ethical approval today.)

When we move to considering policy or practice interventions designed to bring out particular changes, ruling out alternate causes becomes even more difficult. Consider the following scenario.

Two years before the students are due to take their first public exams, a school decides to reduce the size of classes from 35 to 25 because they believe that smaller class sizes will improve exam results. The pass rate for this cohort of students turns out to be better than their predecessors.

Can we infer that reducing the class size to 25 (X) is *responsible* for the improvements in exam results (O), expressed as follows:

Perhaps there were some unusually bright pupils in this cohort, perhaps the school appointed a particularly talented teacher (or a particularly poor one retired), or perhaps something was going on outside the school that impacted on performance *e.g.* a sports club was set up that inspired confidence leading to better school engagement. Let us call this potential extraneous variable (*i.e.* not a part of X), variable Y.



Here, the dotted lines indicate the absence of a causal relationship. In this scenario, the causal relationship is between Y and O, but if we are unaware of Y (or do not take steps to rule out or control its influence) then we might wrongly infer that X causes O.

Threats to valid inference

If someone claims *that X* causes *O*, it is reasonable to ask does the evidence support the claim, and if so to what extent? Is the claim valid?

Validity is a function of the inferences drawn about a causal relationship, rather than a function of any particular research design. Cook and Campbell (1979) identified four types of validity:

- **Statistical Conclusion Validity** – the validity of claims that two variables are correlated or covary one with the other *i.e.* if X then O, and if not X, then not O.
- **Internal Validity** – the validity of a claim that the relationship observed between X (intervention) and O (outcome) is causal.
- **Construct validity** – the validity of claims that what is being done or measured is appropriately described using the higher-order theoretical constructs invoked.
- **External Validity** – the validity of claims that the causal relationship between *X* and *O* will extend to other people, conditions, timepoints.

To accurately determine whether the independent *(X)* and dependent *(O)* variables do, in fact, covary requires the use of appropriate statistical tests (statistical conclusion validity). When we test the effects of an intervention, we typically have (or should have) a theory about what the intervention will involve, how it will bring about the desired effects, and how the effects are most appropriately measured. In light of the results, we consider the implications for our theoretical constructs, which may need to be revisited. Even when we are confident *that X* causes *O*, *if X* is a *complex* social intervention with many components, we may not know which are the 'active' ingredients, so the fact of the causal relationship may have limited explanatory power. These issues are relevant to construct validity. Both statistical conclusion validity and construct validity are important considerations in all research, but in this entry, we focus on those most closely related to study design, namely internal and external validity.

Internal validity

The validity of a causal inference is threatened by anything that suggests O was not caused by X, but by something else which might result in O in the absence of X. Table 1 summarises the list of common categories of threat to internal validity.

Table 1: Threats to internal validity

Threat	Explanation
Selection	Selection bias is perhaps most easily understood as 'allocation bias'. It refers to the systematic differences that can arise between the baseline characteristics of participants in each arm of a trial that can occur in the absence of random assignment, thereby confounding the results.

History	Changes (other than the intervention) may occur during the study period. For example, an intervention designed to help young offenders into work may coincide with a general improvement in employment opportunities; that change, rather than the intervention, may be instrumental in young people finding work.
Maturation	Participants do not go into 'freeze-frame' during a study. They can experience changes which have a bearing on the outcome. Maturing (through age or experience) is an obvious and literal example. Others include getting hungry, bored or tired; feeling more in control of stressors.
Testing	Measurement is necessary, but it itself can affect results. For example, an individual's apparent improvement on a test undertaken for the second time may reflect their familiarity with the test and what is being looked for, rather than a 'true' change that can be attributed to an intervention.
Regression effects	Participants are often recruited to (or volunteer for) studies because they fall into a particular extreme category *e.g.* studies of a therapy for anxiety are likely to recruit individuals who score 'high' on anxiety scales. But extreme categories, by their very constitution, are unstable. Someone starting therapy with high levels of anxiety (a motivator for seeking help) is likely to become less anxious over time, irrespective of the therapy. This is known as 'regression to the mean'. When clinicians talk of the spontaneous remission of psychological problems, they are referring to

	the same phenomenon.
Instrumentation	There may be changes in the calibration of a measure over time; the performance of observers or scorers, or the ways in which data are collected.
Attrition	Not all those who participate in a study complete the intervention and not all participants provide data at all required time points. Few studies will retain all participants or have complete outcome data on everyone. Attrition (also known as 'mortality') is problematic when those who drop out are systematically different from those who remain, or when the reasons for attrition vary between treatment and non-treatment groups.
Selection interactions or additive effects	Threats to validity can operate singly or in combination and can interact in ways that magnify the bias of each. Most commonly a threat exacerbates an existing selection bias, as when non-equivalent groups develop at different rates (selection-maturation) or start off at different points on a measure that is more sensitive to change at certain intervals than others (selection-instrumentation).

External validity

Whilst internal validity is concerned with the internal logic of the research (do our results show what we take them to show), external validity focuses on the external applicability of our results (can we claim that our results generalize to other contexts). Here the adjudication

may be even more fraught, and depend more on background disciplinary knowledge, than in the case of internal validity. Again, a table of issues/threats to consider may be helpful:

Table 2: Threats to external validity

Issue	Explanation
Setting	Can a causal relationship obtained in a factory be obtained in a military camp, or on a university campus? Can an intervention that works for social services in a compact city be equally effective in a rural setting?
Selection	People who volunteer or agree to participate in a study may be systematically different from others in the general population and are almost certainly different from those that refuse to participate. For example, they may be more motivated to change, more altruistic, more impoverished. Thus, the results of a trial may hold only for the unique population from which the experimental and control groups were jointly selected. Will the findings on police intervention still hold if police recruitment or the nature of common crimes change?
Time-point	The observed effect may be dependent on the historical conditions of the time (*e.g.* war, or Covid-19 pandemic, or a time of full-employment). To which periods can a particular causal relationship be generalized?

Unsurprisingly the headings are not unrelated to those for internal validity; many of the issues around generalisability can be re-described as arising from failures to fully specify the causally efficient variables *within* the experiment (so involve issues with internal validity).

But retaining the distinction has heuristic uses. It helps us notice that there may be a tension between internal and external validity. Maximising a study's internal validity can make it difficult to be confident that if the researched intervention is implemented in other settings, by other staff, and with other groups of people, it will deliver comparable results.

Researchers often go to great lengths to ensure that the intervention itself is delivered in a consistent way and to a specified standard, which may include providing training and close supervision of staff. Such 'controls' may not be available outside the experiment. For example, no two organisations are exactly the same, staff groups differ (as do their IT and quality assurance systems), they may service very different catchment areas, access different resources and have very different service user or patient profiles. Further, the outcomes measured in an experiment may not be those of most interest to service providers or users. Will a change in, say, attitude (measured in the experiment) result in a change in behaviour (if not measured) that will persist over time (if not measured)? Designing studies that best minimise threats to internal validity, whilst maximising their external validity is a challenge that is considered below.

Random assignment

Random assignment enables the researcher to establish groups that are probabilistically equivalent. In other words, every participant has the same chance of being allocated to one or other of the study groups *e.g*. a 1 in 2 chance when a coin is tossed; 1 in 6 when a die is rolled. Groups thus established are not only likely to be comparable in relation to those characteristics that we have good reason to believe might influence the outcome of an intervention (*e.g.* IQ, gender, ethnicity, age, class) but on other outcome-influencing characteristics of which we are unaware. Whereas we may be able to 'match' groups on the

former, we cannot ensure that the groups are comparable in respect of characteristics whose influences are important but unknown to us, or that we cannot observe. Further, by definition, randomisation rules out selection bias.

Consider two such randomly selected groups. We can test the impact of an intervention by introducing it to one group only and observing the consequences. Anything else going on for participants in this experimental group is just as likely to be going on in the control group (the untreated group). If an intended (or unintended) outcome occurs only in the group exposed to that intervention then, all other things being equal (which they rarely are), random allocation enhances our confidence that this has been caused by the only systematic difference between them – the intervention. This design – often termed a 'no-treatment control' – is said to have two 'levels', meaning that the causal variable is manipulated so that it is present in one condition (the experimental group) and absent in the other (control group).

Some object to randomisation on ethical grounds; this concern is especially acute when the goal of the experiment is to identify treatments that can produce positive/ desired outcomes. Others point out that in some areas of policy and practice it is neither legal nor logistically possible to randomise participants, whether individuals, groups or communities. History suggests that ethical reasons are often exaggerated, not least because such concerns underplay how little is known about the effectiveness of *current* services and the interventions routinely offered, or the unfairness inherent in their availability. Ethical concerns are usually expressed in relation to withholding a potential beneficial service to those in need, and this is clearly at its sharpest when using a no-treatment control (see below). However, when there is limited resource, randomisation has ethical merit. For example, when a new home visiting service only had capacity to help a certain number of families in a London Borough, randomising eligible parents was judged (by the parents) to be a fairer way of allocating the resource than

leaving it to service managers or professionals.

Common experimental designs

Not all experimental designs are two-group designs, and not all two-group designs compare two groups only one of which is exposed to an intervention. An intervention may be compared with one or more other interventions, or with one or more variations of the same intervention; interventions may, for example, be varied in terms of frequency or intensity ('dose'). Complex interventions may be varied in ways that explore the relative contribution of component parts (sometimes called dismantling studies). **Error! Reference source not found.** Table 3 sets out the structure of commonly used two-group and multiple-group experimental designs:

Table 3: Common experimental designs

i	Pre-test Post-test Control	O_1	R	X	O_2
		O_3	R		O_4
ii	Solomon four group design	R	O_1	X_A	O_2
		R	O_3		O_4
		R		X_B	O_5
		R			O_6
iii	Post-test Control		R	X	O_1
			R		O_2
iv	Post-test – two interventions*		R	X_A	O_1
			R	X_B	O_2
v	Pre-test Post-test – two interventions*	O_1	R	X_A	O_2
		O_3	R	X_B	O_4

*These designs can have more than two intervention groups, and may also include a control group (not shown)

Each row shows the handling of one group in the experiment, using the following common code:

X represents the exposure of a group to an intervention (the *experimental* variable, also known as the *independent* variable);

O represents the observations or measurements taken (also referred to as the *dependent* variables);

R represents randomisation, conceived as 'the all-purpose procedure for achieving pre-treatment equality of groups, within known statistical limits'(Campbell & Stanley, 1963 p.6).

The three most commonly used designs are introduced here:

Pre-test post-test control group designs

The classic controlled experimental design is the pre-test post-test control group:

Treatment/intervention group:	O_1	R	X	O_2
Control group:	O_3	R		O_4

In this design, the effect of *X* is determined by comparing the difference in the post-test and pre-test measures between the two groups. The effect is the difference between $(O_2 - O_1)$ and $(O_4 - O_3)$. We have here shown O_1, O_3 as *preceding* R; logically when both groups are pre-tested this sequence is immaterial, but in practice it may be more robust to apply pre-testing to the whole group before any randomisation steps are taken,

The strength of this design lies in its ability to control for almost all *threats to internal validity*. Taking each of the these in turn, beginning with history.

History Both groups are equally exposed to changes in external context (but the researcher has to remain alert to any events – other than X – which affect one group but not the other).

Maturation and testing Both groups experience the same tests at the same times, thereby controlling for *maturation* and *testing*.

Instrumentation Controlling for *instrumentation* can be challenging if we move from standardised measurement instruments completed by participants. Observers or interviewers may 'frame' treated and control participants differently. One defence is to conceal the experimental status of participants from those collecting data; this is not always easy to achieve, not least because participants often inadvertently disclose their status by referring to the intervention received.

Regression to the mean Randomisation ensures that both groups can be expected to regress equally. The challenge for researchers is not to be seduced into over-interpreting (*mis*interpreting) the changes made in any subgroup of participants who may have had more extreme scores at pre-test. Looking at the changes made by those with extreme scores in both groups can provide a counterbalance.

Randomisation also controls for *selection bias*, as long as the sample size is adequate. Unless subverted (and history provides evidences that subversion is sometimes attempted), randomisation prevents systematic differences arising between groups that might confound the results of the trial. Nevertheless, in order to check for chance differences between experimental and control groups it is good practice, particularly with smaller samples, to collect attribute data on variables plausibly related to our outcomes (age, sex and race are always likely individual-level contenders; informed reflection on the particular outcome will suggest others). The analyses can then adjust for these.

Attrition is a significant problem throughout social science methodology. Sample surveys can

have low response rates, and panel studies suffer drop-out. In experimental designs, randomisation is designed to establish equivalent groups, but attrition can mean that the groups at the end of an experiment are no longer equivalent. Attrition reduces statistical power and – if participants are lost from each group for systematically different reasons – may threaten the experiment's internal validity. There is much the active researcher can do to minimise attrition (see Greenberg & Barnow, 2014; Ribisl et al., 1996), and decisions regarding how to deal with attrition at the data-analysis stage is a key consideration.

Competitor explanations for the causal inference from *X* to *O* are not the only problem facing the experimental researcher. In some contexts, interaction effects between *X* and another variable within the experiment pose a particular threat to the experiment's *external validity*. For example, in an experiment designed to assess the effectiveness of a television drama aiming to change attitudes to alcoholic misuse, a pre-test of viewers' attitudes towards alcohol might 'sensitise' them to the key messages in the drama (thereby enhancing its effect). In some cases, a pre-test might be so influential that one might not be confident that an intervention would have an effect in its absence. Alternatively, it might blunt the message conveyed by the drama. The significance of such an interaction will vary with the context and how 'out-of-the-ordinary' the content of the test might be for those tested. In these circumstances, designs with no pre-test might be preferable; the Solomon four group design (which we will shortly discuss) can test for the magnitude of the interaction.

We have noted that randomisation deals with the possible biases that would arise from *selection* into 'treatment' and 'no-treatment' groups. There remains the issue of selection into the experimental process in the first place. The interaction between the intervention and such selection is probably a more serious threat to external validity, concerned as it is with the ways in which those involved in an experiment may differ from those who are not. This can

be an artefact of whom the researcher approaches or who agrees to participate in an experiment (those who agree may differ systematically from those who do not). Protocols for random sampling can minimise this, but an alternative is to purposively sample in ways that maximise heterogeneity, either of individuals or settings.

Logically, of course, we only ever make a *judgement* that the results of an experiment will generalise to other groups and contexts; we return to this issue below.

One design that explicitly addresses issues of external validity in relation to the 'testing' environment of experiments is the Solomon Four Group Design.

Solomon Four Group Design

In this design participants are randomised to one of four groups that effectively pair a post-test only design (rows 3 and 4) with a pre-test post-test design (rows 1 and 2). This design enables the main effects of testing and the interaction effect of testing and the intervention to be ascertained. The researcher can identify:

1. Whether *pre-testing* has influenced the results, by comparing the difference between the results of O_5 and O_6 and O_2 and O_4.

R_A	O_1	X_A	O_2
R_B	O_3		O_4
R_C		X_B	O_5
R_D			O_6

2. Whether any *external factors* during the experiment have influenced the outcome, by comparing O_3 (the group pre-test of R_B) with O_6 (the post-test results of R_D).

R_A	O_1	X_A	O_2
R_B	O_3		O_4
R_C		X_B	O_5
R_D			O_6

3. The effect of the pre-test on the treatment, by comparing O_2 and O_5. Any difference identified would indicate that the pre-test had influenced the intervention.

R_A	O_1	X_A	O_2
R_B	O_3		O_4
R_C		X_B	O_5
R_D			O_6

4. Whether the pre-test itself influenced behaviour independent of the intervention, by comparing O_4 and O_6

R_A	O_1	X_A	O_2
R_B	O_3		O_4
R_C		X_B	O_5
R_D			O_6

The Solomon Four Group also makes a useful, albeit indirect contribution to issues of the generalisability of findings emerging from pre-test post-test control group designs. In any given area, the findings of Solomon Four Group designs tell us about the likelihood of interactions between tests and outcome, thereby helping with the interpreting of findings that do not control for this; similarly, with regards to the combined effect of maturation and history. Whilst a very powerful design it is quite demanding in terms of implementation, and represents a big ask of potential participants, particularly organisations.

Post-test control group designs

As mentioned, the Solomon Four Group Design allows estimation of the magnitude of any pre-test effect (panels 1 and 4 of the previous section) and the pre-test's interaction with the intervention (panel 3). In contrast, the 'post-test only' design handles the pre-test problem by the simple expedient of omitting any pre-test.

R	X	O_1
R		O_2

This design tends to be eschewed, due to concerns that randomisation does not always result in groups that are equivalent on variables known to be important. Pretesting can provide important information about participants, which can help to determine whether those who drop out, or for whom data are not available, differ systematically from those who do not. This does not, however, undermine the importance of randomisation in establishing probabilistically equivalent groups.

The post-test only design may, however, may be the design of choice in certain circumstances. In some contexts, testing is unnecessary since relevant baseline data are already available *e.g.* numbers of children achieving a certain grade at age 10 for mathematics; the numbers of people claiming to be in work. For some problems a pre-test may not be possible. If we want to study the impact of nurse home-visiting from 20-weeks' gestation on the cognitive development of infants subsequently born, then it is not possible to do a 'pre-test' for this particular outcome. Sometimes pretesting will be prohibitively costly. And of course, there may be reason to believe that it will influence the outcome or interact with the intervention.

These designs convey the basic principles underpinning experimental research. **Error! Reference source not found.** sets out three more designs.

Table 4:Three further designs

Longitudinal designs	R	O ... O	X		O	O … O
	R	O ... O			O	O ... O
Cross-over designs	R	O	X_A	O	X_B	O
	R	O	X_B	O	X_B	O
Factorial designs	R		X_{A1B1}		O	
	R		X_{A1B2}		O	
	R		X_{A2B1}		O	
	R		X_{A2B2}		O	

In randomised *longitudinal* designs, pre-test data are collected at multiple point before and after the intervention, or simply at multiple points after the intervention. These designs enable researchers to examine the impact of an intervention over time, although attrition can be a particular problem. *Crossover* designs are experiments in which participants are randomly assigned to receive two or more interventions in a predetermined order. Observations are made after each exposure and may be made pretreatment. They are common in medical research where, for example, researchers can organise a 'wash-out' period between exposures to each 'treatment' (usually a drug). In social science research it is more difficult to 'undo' the impact of the first 'exposure', which then carries over to subsequent exposures, acting as a 'confounder'.

Factorial designs enable us to investigate a range of factors that might influence the outcome of an intervention. In these designs, participants are randomly assigned to all possible combinations of factors under investigation (independent variables), typically at different 'levels' (a subdivision of a 'factor'). For example, we might be interested in whether or not a certain kind of phonics tuition works best when delivered on-line or in person, and whether it is more effective when children receive hour-long lessons three times a week, or ninety-

minute lessons twice a week. Participants would, in this design (called a 2x2 factorial design) be randomly allocated to one of four conditions, as illustrated in Table 5.

Table 5: A factorial design example

			Factor B	
			Level 1	*Level 2*
			3 × 1 hour	2 × 1.5 hour
Factor A	*Level 1*	Online	Group A_1B_1	Group A_1B_2
	Level 2	In person	Group A_2B_1	Group A_2B_2

Each group can then be compared with the others. The number of factors and levels can be increased to address more complex questions. The advantages of factorial designs include their efficiency (they often require fewer participants or units to be randomised), their ability to shed light on which combination of factors is most effective, or simply whether a combination of factors is more effective than one, and their ability to test interactions. These designs require particular approaches to their analyses that differ from simple RCTs. For a full discussion and elaboration of these issues see Linda Collins (2018) in the further reading.

Key considerations in experimental design

The apparent simplicity of a randomised experiment is beguiling. The realities of designing an experiment in the field can, however, be very complex and challenging, with many technical issues to consider once the basic design has been chosen. Effective implementation also requires close attention to practical issues, such as how to work effectively with collaborators, how to engage potential participants, and how to retain the interest and commitment of all parties during what can feel a long time. The successful delivery of a real-

world experiment requires implementation skills – often from a range of disciplines, effective project management, and good interpersonal skills. It is essentially a team affair. In this section we consider some of the key decisions in the design of an experimental study, rather than in its implementation. First, some ethical considerations, and the role of theory.

The ethics of randomisation

Arguably, it is unethical to involve participants in a study designed to answer a question that has already been answered, or one that is not sufficiently well designed to be able to answer the question it addresses. For these reasons, some funders will not fund primary research until applicants can demonstrate that they have conducted a comprehensive review of current best evidence, typically in the form of a systematic review. As with all research, experimental designs should not put participants at risk and should meet the same ethical requirements as all research.

Some argue that, because of randomisation, experimental designs are intrinsically less ethical than others. However, as we have already noted, the history of social intervention (by governments, professionals, and others) makes clear that what seems intuitively or theoretical obvious may fail on systematic inspection. For example, it is not unreasonable to think that one might be able to deter children at risk of delinquency from a lifetime of crime by bringing them face to face with imprisoned offenders who could give them vivid, yet realistic, accounts of life in prison. The underpinning rationale seems plausible. The young may well not have an accurate understanding of the lifetime costs of criminal activity; meeting prisoners might well provide a powerful ‘reality check’. Accordingly, programmes have been designed to ‘scare’ young people into change, and the best known is ‘Scared Straight’. Early observational and quasi-experimental studies of Scared Straight produced a

mixed picture of impact, but most suggested it was effective. Later, experimental studies firmly pointed in the opposite direction, namely that those involved in Scared Straight, were more likely to be involved in delinquency compared with youths who were not exposed to such programmes (Petrosino, Turpin-Petrosino, Hollis-Peel, & Lavenberg, 2013). These negative results also applied to alternate programmes designed to be more educational than frightening.

Leveraging theory

The focus of most policy experiments is finding out what works (or doesn't), what works better or what works most cost-effectively. Answering these questions depends on having a sound understanding of the nature of the social issues that we are seeking to influence, the contexts in which they occur, and their impact. Knowing what has been tried before, with what results, and the likely reasons for these, are important 'upstream' questions.

Developing and articulating clear theories as to why doing '*this'* will result in '*that'* is important. Such theories assist programme developers to investigate interventions that are most likely to have the desired impact. They also help researchers to design experiments in ways that maximise interpretability of the results. However, as anyone who has tried to develop a theory of change knows, it can be challenging. There is no agreed prescription for how to go about developing, or using, a theory of change, but consideration needs to given to the following:

Context for change – consider all factors of potential relevance. Even the simplest social intervention occurs within a complex social system, for example, an organisation or professional team, in turn nested within broader 'supra-systems' (Moore et al., 2019).

Identifying potentially relevant factors at all levels, from the micro to the macro, are important considerations in thinking about the 'pathways to impact', bearing in mind also that few organisations 'stand still' throughout an evaluation.

Desired outcomes – who is intended to benefit from the intervention and in what ways.

Process of change – an account of how change will be achieved, and the assumed relationships between component parts. Detailing the sequence of events that is thought to result in a particular outcome or set of outcomes is at the heart of a theory of change. Too often these sequences are represented as simple linear models of causality without due attention to the inherent complexity of change endeavours.

Assumptions– identification of the assumptions underpinning each step or 'link' in the presumed process of change. Assumptions should be provisional, testable and revisable. A systems perspective would suggest that it is sensible to consider the range of things that might happen and that might impact on what actually happens, as opposed to what we have planned.

Theories of change benefit from the involvement of all key stakeholder groups, including the intended beneficiaries (where possible), those delivering the intervention, their managers and any other key figures within the organisational context. Such involvement is important in ensuring that programme developers and researchers have an adequate understanding of context and can help steer the development of interventions towards the acceptable and deliverable. They can also help to identify ways in which the implementation of a programme might fail, with implications for the evaluation approach.

These elements are usually brought together in a summary statement and may be depicted

schematically, though this is not essential; the acronym PICO(T) has been adopted by experimental researchers.

PICO(T) and asking an answerable question

Experimental designs test hypotheses about the effects of an event or intervention in a particular set of participants, usually in a particular context, and with expectations (hypotheses) about the impact of that intervention. It is generally accepted that in order to frame an answerable question, we have to ensure the following are clearly identified: the problem or population of interest (P), the intervention to be studied (I), the comparator (C), the outcomes to be evaluated (O) and, if relevant, the time duration for either the intervention or the point of measuring the outcome, or both (T): PICO(T).

The P in PICO is a key factor in determining the inclusion criteria for the study. From which population will participants be recruited? How will we define who is 'at risk of delinquency'? Will we recruit only those at high risk and how will these be defined? Will we include males and females, or are there reasons to focus only on one group? What age range? From one location or more than one? Do we exclude those with other difficulties? The answers to these, and other questions about 'who' or 'what', facilitates explicitness and transparency, both of which are key to developing an evidence-based applied social science.

The I in PICO highlights the importance specifying and delivering the intervention 'as intended'. This concern with 'treatment fidelity' may point to the need to extensively and formally document ('manualise') an intervention and provide training and supervision to those delivering it. A manualized intervention can be designed to support flexible delivery, but recording how it is implemented becomes particularly important.

The O draws attention to what it is that the intervention is expected to achieve. What difference do we expect it to make, and how best is that assessed? Sometimes the outcomes of interest are longer term (distal), though researchers often choose to measure proxy outcomes that are believed to act as intermediary indicators signaling the likelihood of those distal changes. Sometimes we can follow up participants long enough to track those distal outcomes, but this possibility is often curtailed by funding constraints.

The choice of comparator (C). A key challenge in designing experimental studies is to provide a 'fair test' of an intervention's effectiveness. As well as giving due consideration to what an appropriate control would be, researchers should ensure that they accurately describe the services available to, or accessed by, control participants. This is necessary to understand the reasons for any differences observed between intervention and control group, and the resulting generalisability of the findings. Although considerable attention is now paid to *treatment* fidelity, it is rare to find similar attention paid to interventions or services available to participants in control arms. This can result in threats to the internal validity of the causal inference. For example, can we be sure that control group participants received 'nothing' in the no-treatment arm of a trial, or are we clear what 'management-as-usual' actually involved? Where professionals are, under treatment-as-usual, deploying an intervention that they are confident will deliver the intended benefits, this may impact on how they perform, and on the outcomes; it may also lead them to be lax in the implementation of the *trialed* intervention as new-fangled frippery. Where professionals are overworked, poorly supervised, or dispirited in their efficacy, this may impact on how they perform, and on the outcomes; it may also lead them to enthusiastic support for the trialed intervention simply because it is *not* business-as-usual. These last concerns highlight the importance of collecting information on the process and implementation features of studies.

Which experimental design?

When evaluating a novel intervention, there may appear to be a strong argument for using a no-treatment control or a wait-list control (in which the control group get the intervention after the impact of the intervention has been assessed for the experimental group). Whilst these designs can provide clear evidence of the effect of an intervention, and evidence of what happens in its absence, they are not always possible or appropriate, sometimes for ethical reasons and sometimes for purely logical ones. For example, circumstances may dictate that all participants *must* receive *an* intervention, as when all children must be taught math (legal requirement), or those requiring medical care should receive it (ethical and possibly legal requirement). A wait-list control may reduce the apparent inequity (in the long run, all participants receive the treatment) but for many treatments, time matters; for patients and for pupils, tomorrow differs from today. In such circumstances, researchers need to find alternative means of controlling for threats to internal validity.

Available options for control conditions include management-as-usual/treatment-as-usual; non-specific component controls and specific component controls. As always, the driver should be our theoretical and substantive concerns. Suppose a researcher seeks to evaluate 'photo reminisce' therapy for Alzheimer's sufferers, in which the therapist spends an afternoon exploring family photographs with the participant. Is the appropriate contrast with management-as-usual (yet another afternoon sitting isolated in the residents' lounge), or an afternoon of conversation (not photo-based) with an engaged therapist? If the latter, how can that condition be specified? The choice of control will depend on the question one is seeking to answer, and each has its own advantages and disadvantages.

- Management-as-usual / treatment-as-usual controls. Here, a new intervention is compared with those already in use in a particular setting or organisation and provides information about whether or not it is more effective, equally effective, or (where appropriate data are available) better value.
- Non-specific component controls. In these designs, researchers separate out the specific 'ingredients' of an intervention *e.g.* cognitive behavioural therapy (CBT) from the non-specific factors which might act as a catalyst for improvement *e.g.* the therapeutic relationship, the clinical or professional setting, the attention given to the participant.
- Specific component (or dismantling) controls. In these designs, researchers try to evaluate the particular contribution of parts of multicomponent or complex interventions. Most real-world social or psychological interventions are multicomponent and complex, so these designs should have much traction. In these designs the control group receives some, but not all, of the component parts of the intervention. For example, when researchers looked at treatment of phobias and manipulated the components of *systematic desensitisation* (relaxation, developing a hierarchy, imaginal exposure and *in vivo* exposure) to determine which were necessary and sufficient, the results led to the deployment (for many people at least) of a briefer therapy, namely graded *in vivo* exposure.

When and how to randomise?

Participants in an experiment can be individuals, small groups or units (*e.g.* classrooms, teams), or larger units such as communities or villages/neighbourhoods. The latter are generally referred to as place-based or cluster randomised trials. The choice of unit of randomisation is considered in the next section. This one considers the mechanics of

randomisation.

Because of its role in minimising bias, it is important that the method used will guarantee that assignment is ‘by chance’. There is ample evidence that service providers with strong views about an intervention or about randomisation can subvert randomisation, as can investigators, concerned that simple techniques (such as coin tossing) are resulting in groups of uneven size. As Mosteller and colleagues observed ‘When the randomization leaks, the trial’s guarantee of lack of bias runs down the drain’ (Mosteller, Gilbert, & McPeek, 1980 p.40). We consider procedures for avoiding uneven group size later, but we begin with a discussion of approaches that can guarantee ‘allocation by chance’.

The most usual method now used is a random allocation numbers table, in which a computer generates a table of random numbers in some specified range, for example, integers in the range 0 to 99 as in Table 6. Having generated the table, the responsible statistician (or team member) establishes a decision-making rule covering where to start and what direction to read the numbers.

Table 6: Example of a random numbers table

26	76	8	72	37
33	63	27	12	46
65	18	07	55	36
51	24	52	19	755
28	79	67	43	77
35	15	78	25	59
09	38	48	62	20
10	44	80	01	02
47	41	13	40	42
50	68	61	70	6

For example, in a two-arm trial, we might allocate odd numbers to group A and even numbers to group B. Thus, in Table 6, with number 24 in column 2 as the start point, we would have the following sequential allocation of participants to group A or B: ABBAABAABB. There are other rules that use grouping of numbers *e.g.* 00 to 49 for A and 50 to 99 for B; or using each digit on its own, and the approach can be extended to cover trials with more than two arms.

When sample sizes are small (say under100), in any *particular* sample randomisation will not necessarily result in groups of equal size and may display between-group differences in characteristics related to the outcome (covariates). For example, in an experiment to evaluate a scheme aimed at getting unemployed people into work, length of time unemployed and severity of mental health problems might make getting back to work more difficult. To devise

a ‘fair test’ of the intervention we need to avoid having participants with these characteristics over-represented in either group. Randomisation can be restricted in ways that help overcome these problems.

Block randomisation Groups of equal size can be achieved by first grouping participants in ‘runs’ or blocks of allocations and then randomising from within that block. Consider a run or block of four allocations: there are six different ways of arranging allocation to A and B within such a block that ensures that we add exactly two participants to group A and two to group B. These are (in no particular order): AABB, BBAA, ABAB, BABA, BAAB, ABBA. We then randomly choose *one* of these allocation schemes and use it to allocate our first four participants; make another random choice of scheme for our next four, and so on. This mechanism ensures that, wherever we stop the allocation, the numbers selected into A and B will never differ by more than half of the length of the block (in this example, two cases). This approach does not constrain other aspects of the randomisation.

Stratified block randomisation This method of block randomisation can help to avoid between group variation on a small number of covariates. Participants are first grouped according to the covariates of interest *e.g.* length of time unemployed, severity of mental illness. Block randomisation is then used to allocate participants from within each group to the arms of the trial.

Because of the predictability inherent within each block, responsibility for allocation is usually restricted to the statistician on the team or otherwise masked from the research team. Using blocks of different sizes can also help to minimise predictability. This approach is only of use when researchers have identified all participants prior to randomisation. This makes it difficult to deploy in experiments that recruit participants over a long period of time.

As implied above, the most common allocation ratio in a two-group trial is equal numbers in each arm. It is, however, possible to allocate on a different ratio. Practical or ethical constraints may entail unequal arms, as in an evaluation of a group-based version of Nurse Family Partnership (Barnes et al., 2017). Recruitment to the study was challenging, and the research team required a minimum of eight couples to form a group in the experimental arm. To achieve this, and after careful consideration of the impact on the study's power, the team opted for a ratio of 2 to 1 in favour of the intervention group. In some circumstances, unequal sample sizes can increase power, but these considerations are beyond the remit of the present introduction.

Units of randomisation

Consider an experimental study to evaluate a new approach to training professionals in risk assessment. Should we randomise the professionals, the teams they work in, or the organisations that employ them? These are questions about the unit of randomisation, which has been described as simply 'an opportunity to apply or withhold the treatment' (Rosenbaum, 2010, p.23). Some studies randomise those delivering an intervention (*e.g.* doctors, counsellors, teachers). Some randomise naturally occurring groups (*e.g.* families, classrooms, year groups, teams of workers) or those brought together specifically for the purposes of the study (*e.g.* support groups, psychoeducation classes). Places can be randomised (*e.g.* streets, neighbourhoods, regions, fields), and so too can the timing of intervention, as in single case experimental designs. In industry, the unit might be batches of raw material (Cox, 1961).

The choice of unit of randomisation will, in part, be driven by the focus of the intervention. Thus, an evaluation of systemic family therapy (SFT) is likely to point to the randomisation

of either families (some to receive SFT, others not) or of therapists working with families (some delivering SFT, others not). As with other aspects of experimental design, there are practical, financial, ethical and scientific considerations that have to be taken into account in determining the appropriate unit.

A key scientific consideration is the avoidance of contamination or spillover effects from the diffusion of treatments. The purpose of randomisation is to ensure there is 'clear blue water' between those in the intervention group and those in the control group, such that the experiment reveals what would happen in the absence of the intervention. This is referred to as the 'independence principle'. A study in which spillover effects result in a breach of the independence principle is flawed, and often these effects are avoidable with better 'upstream' planning.

For example, Yokum, Ravishankar, and Coppock (2017) studied the impact of Body Worn Cameras (BWCs) on the use of force by police, and on civilian complaints. A key theory of change behind this intervention was that police officers and the public might be expected to behave differently when being watched: it was expected that BWCs would make it more likely that police would adhere to departmental protocols (particularly regarding unjustified use of force), and that civilians would be less likely to behave badly:

> The underlying pathways between BWCs and behavior could include greater self-awareness, heightened threat *of being caught, or a combination of the two. (Yokum et al., 2017 p.2)*

In all, 2,224 individual police in seven districts were randomised to either the experimental (BWCs) or the control (no BWCs) group, and the impact assessed at seven months. The sample size was large enough to detect even a small impact, but no differences were found

between the two groups. This could have meant that BWCs had no effect on police behaviour, for whatever reason, but it is more likely that no difference was found because of spillover effects. Police officers often work in pairs and, when working alone, one officer without a BWC may respond to an incident alongside another wearing one and behave differently as a consequence. The authors found that in 30% of calls for services, officers from *both* arms of the trial responded. In short, the assumption of 'independence' between the two arms of the trial was violated, with the impact of police officers wearing BWCs present in both groups. Arguably, careful discussions with key stakeholders could have highlighted these implementation challenges, and alternate – less problematic - units of randomisation selected.

The BWC study is an example of spillover effects when the unit of randomisation is the individual; when this risk is evident, it points to the need to consider randomising at a higher level. How much higher depends on the nature of the intervention and the context. The challenge is to choose a unit of randomisation in which we can be reasonably confident that the units will not influence one another. When we can minimise, but not eliminate, spillover effects, we can take steps to take account of it, either at the design phase (for example, by including additional control groups that are external to the trial) or, more commonly, in the analyses of the data.

Implications of cluster randomisation

Whilst there are often sound reasons for choosing to randomise higher level units, rather than individuals, certain costs follow. Randomising clusters of individuals affects sample size calculations, analyses, and reporting.

Put simply, individuals in a cluster are more likely to be similar to one another in important respects than individuals who are not recruited from a cluster. For example, children in a class are likely to be similar given their shared exposure to the same teacher and their ways of teaching and learning. Some schools organise classes by ability. Children interact with one another, and therefore influence one another. At the school level, schools recruit from particular geographical areas that vary in significant ways. Teachers *within* a school may share knowledge, strategies, attitudes and ways of working with others in the school. Between-cluster variation is also something that needs to be considered.

Cluster randomisation is said to be statistically 'less efficient' because the information obtained is less than if the same number of *individuals* were individually randomised. Participants in a cluster are likely to respond to an intervention in similar ways (for a variety of reasons) and so are statistically less 'independent' of one another. Observations of impact at the level of the individual will therefore be correlated. The extent of similarity *within* a cluster is measured by the parameter referred to as the intraclass ('intra-cluster') correlation coefficient (ICC). Problems arise in cluster randomised trials when the data are analysed at the level of the individual without taking account of the diminished independence of observations *within* each cluster *i.e.* without using the ICC (which may be known or may have to be estimated). Most textbooks on cluster randomised trials refer to the work of Jerome Cornfield who, in 1978, drew attention the particular statistical features of cluster randomisation, and cautioned that 'randomization by cluster accompanied by an analysis appropriate to randomization by individual is an exercise in self-deception' (p. 101-102). Because participants in a cluster randomised trial are no longer independent individuals, the analyses have to take into account both within-cluster and between-cluster variation. The effective sample size is reduced, and statistical power is accordingly reduced, resulting in

larger standard errors and less precise estimates of the impact of the intervention (wider confidence intervals). The fewer clusters there are, and the larger the cluster size, the greater these impacts, and the more likely it is that results will be misleading unless clustering is adequately taken into account.

Sample size

We should not expend more resource or involve more people than is necessary to identify a causal relationship or falsify the null hypothesis (no effect of the intervention). But to proceed with a study that had *too few* units to identify a relationship would be bad science. Bad because inefficient and likely to waste resources, and bad because it can mislead. To report 'we found no effect' (when our study lacks the power to find effects) may lead readers erroneously to believe that there is in reality no effect. How then do we decide on the optimum sample size for a study? Too small a study means that those participating in it are participating in a study unable to answer the question it is designed to address. Identifying the sample size needed to detect an effect of an intervention (beneficial or detrimental) is therefore an important ethical, as well as a scientific issue.

In determining the sample size required to determine whether the intervention and effect covary, we draw on the concept of statistical power. Formally, the power of an RCT measures how likely the study is to produce a statistically significant result, assuming a population difference of an assumed magnitude – we are worrying about our ability to detect true positives. Characteristically we *i*) decide on an appropriate power level, *ii*) specify various substantive features of our data (such as the normal outcome variation, the size of effect expected), and *iii*) 'calculate the sample size. Specifying the formulae to derive sample size is not an appropriate task for this entry (these formulae evolve, are complex, vary by

research design, and sometimes simulation would anyway be a better route). But the first two steps are the province of the substantive researcher, who needs to provide the statistician with relevant information about the proposed population, intervention and outcomes to be assessed, and we conclude section be examining these in more detail.

Appropriate power levels

The target numeric value for statistical power is not writ in stone, but convention has settled on 80%. This convention regards it as reasonable to embark on an RCT if it is large enough to stand an 80% chance of detecting an existent effect. In other words, it is acceptable to have a sample size such that in one out of every five times such a trial was conducted, we would *fail* to find an effect which does in fact exist.

Of course, any researcher with a low powered design may strike lucky and detect an effect, but claiming luck, of this nature, is not a good claim on research funding. Perhaps more seriously, any positive finding from an underpowered study is likely to *overestimate* the true effect. This, coupled with publication bias towards positive results, can impart a marked upwards skew to *reported* effect sizes.

'Need to know' features for estimating sample sizes

Suppose, following the currently accepted convention for 'good' practice, we opt for a study with 80% power. In order to inform the sample size calculation, the researcher needs to draw on the extant literature, theory, or informed intuition, to make estimates of the size of the expected effect of the planned intervention. If the reasonable expected effect is subtle, we may need many cases to detect it; if large, we can get by with fewer cases; but in reading that literature the researcher should be alert to the upward biases mentioned above.

Next, we need to decide on a target significance level for results; again, convention has settled on a 5% level, but convention may not be appropriate for all investigations. Remember also that if we obtain a particular effect, estimated to be significant at the 5% level, that tells us there is a mere 5% chance of such a finding in a world in which there is really *no* effect; it is not an endorsement of the *particular* value estimated (whence it is good practice always to report a confidence interval for the estimate).

The final piece of information the researcher should routinely provide to the statistician is the natural variation of the outcome measure. Intuitively, intervention effects are more difficult to detect (and need larger sample sizes) if the outcome is naturally highly variable.

Further considerations

As we move away from the basic RCT design in which individuals are the unit of randomisation, additional information may be required. For example, as discussed above, if we deploy cluster randomised designs, different sizes of cluster or different ICCs (within cluster variation) affect the power of an experiment using this design. Other design variations each present their own issues for power calculations. Again, the advice is: talk to your colleague statistician. Knowing the impact on power, we are then able to reflect on the research design trade-offs that present themselves.

One further social science issue is worth mentioning. Even quite advanced textbooks (perhaps because anchored in 'hard' medical practice) pay little attention to error in outcome *measurement*. Suppose we are interested in whether a certain psychosocial intervention attenuates depression. There are well validated questionnaire scales measuring depression. But, however well validated as measures, these are still liable to false positives and false

negatives. Since we are interested in detecting the impact of our intervention on *depression*, not on the *scale* measuring depression, consideration of the prevalence of these errors should enter into our power calculations.

We have been discussing sample *size*. Another important resource issue is choice of sample *origin*. For the researcher, certain geographic areas, for example, may be easier to target than others – perhaps simply because closer, or because the service administrators therein are known to be sympathetic to research. We met this issue when discussing the P in PICOT. Findings from any sample can only be generalised to the population from which they were drawn, though the seriousness of this constraint may be open to informed substantive discussion. For certain interventions, efficacy might, arguably, be largely unaffected by a research focus on only one particular geographic region. For other interventions their observed performance may depend crucially on the idiosyncratic features (cultural, institutional, economic, ethnic) of that specific region. The researcher must take care over claims to generalisability (external validity), and, in deciding where to sample, give thought to what generalisations are substantively desired.

Reporting and quality appraisal

Assessing the internal validity of a trial depends on transparent reporting, yet experimental studies are often poorly reported. There is now detailed guidance available on how researchers should report experimental studies known as the CONSORT Statement (Schulz, Altman, & Moher, 2010). CONSORT stands for Consolidated Standards of Reporting Trials, and the CONSORT Statement comprises a 25-item checklist and a flow diagram, designed to ensure that researchers report on all key aspects of a trial. It is accompanied by a more detailed 'explanation and elaboration' document, which provides an explanation of why each

item is judged to be important and provides additional guidance. Extensions to the CONSORT statement provide guidance for other experimental designs. Whilst motivated by concerns over reporting, the Statement and its Extensions also provide useful guidance to those engaged in developing study designs.

Critiques

Early critiques of experimental strategies in social science tended to focus on their ethical dangers. In the eighteenth century, Adam Ferguson, noting the utility of experiments in physical science, remarked:

> *"[Experiment] is a method of observation which cannot be equally pursued in the study of … human affairs … No man is so much the master of his fellow creature, as to claim the right of exposing them to the risk of a trial" (Ferguson, 1792 vol 1, p.96).*

Such issues were discussed above. There is also an active literature advancing a stronger critique, arguing that an emphasis on RCTs as the 'gold-standard' of social science evidence is misplaced. Part of this critique is simply empirical, pointing out that (some) treatment effects have failed to generalise to fresh situations in the way that enthusiasts might have hoped. Part of the critique derives from a more articulated philosophical position. Its main narrative runs:

- To focus merely on 'what works' is an *a-theoretical* endeavour; without an explanatory *theory* we have no grounds for assuming that an intervention that works for population P_1 at time T_1 will work for P_2 or work at T_2.
- RCTs are not a good tool for theory generation.

- This absence of theory is then posited as the ground of the empirical failings noted.

The point about the importance of explanatory theory is well taken. It applies equally to observational and experimental studies. We can indeed only move beyond the observed if we have an accurate model of the causal processes generating the observed.

The depiction of the theoretical disutility of experimental designs is a more contestable claim. Remember that throughout this introduction we have had to emphasise the importance of articulated theory in designing experiments, and in understanding the import of our conclusions. Only by being anchored in theory can we decide whether, and to where and what, our findings might generalise. But – and it is a non-trivial caveat – unless we have the arrogance of a Newton ('*Hypotheses non fingo*') our theories inevitably must remain our accounts of reality, and not reality. Experimental designs remain good tools to ascertain whether our theoretical accounts fit reality. And the strategy of randomisation enables us to control for these variables and causes which our best theory may have overlooked.

Further Readings

Collins, L. M. (2018). *Optimization of behavioral, biobehavioral, and biomedical interventions.* New York: Springer.

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

Glennerster, R., & Takavarasha, K. (2013). *Running Randomised Evaluations: A Practice*

Guide. Princeton, NJ: Princeton University Press.

Hayes, J. R., & Moulton, L. H. (2009). *Cluster Randomised Trials.* Boca Raton, FL: Chapman & Hall/CRC.

Rossi, P. H., Lipsey, M. W., & Henry, G. T. (2019) *Evaluation: A Systematic Approach.* (8th Edition). Thousand Oaks, CA: Sage.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Boston: Houghton Mifflin.

References

Barnes, J., Stuart, J., Allen, E., Petrou, S., Sturgess, J., Barlow, J., . . . Melhuish, E. (2017). Results of the First Steps study: a randomised controlled trial and economic evaluation of the Group Family Nurse Partnership (gFNP) programme compared with usual care in improving outcomes for high-risk mothers and their children and preventing abuse.

Campbell, D., & Stanley, J. (1963). *Experimental and Quasi-experimental Designs for Research*. Boston: Houghton Mifflin Co.

Cornfield, J. (1978). Randomization by group: a formal analysis. *American Journal of Epidemiology, 108*(2), 2.

Cox, D. R. (1961). Design of experiments: The control of error. *Journal of the Royal Statistical Society. Series A (General)*, 44-48.

Ferguson, A. (1792). Principles of moral and political science: Being chiefly a retrospect of lectures delivered in the college of edinburgh. In two volumes (Vol. 2): A. Strahan and T. Cadell, and W. Creech.

Greenberg, D., & Barnow, B. S. (2014). Flaws in evaluations of social programs: Illustrations from randomized controlled trials. *Evaluation review, 38*(5), 359-387.

Moore, G. F., Evans, R. E., Hawkins, J., Littlecott, H., Melendez-Torres, G., Bonell, C., & Murphy, S. (2019). From complex social interventions to interventions in complex social systems: future directions and unresolved questions for intervention development and evaluation. *Evaluation, 25*(1), 23-45.

Mosteller, F., Gilbert, J. P., & McPeek, B. (1980). Reporting standards and research strategies for controlled trials: agenda for the editor. *Controlled Clinical Trials, 1*(1), 37-58.

Petrosino, A., Turpin-Petrosino, C., Hollis-Peel, M. E., & Lavenberg, J. G. (2013). 'Scared Straight'and other juvenile awareness programs for preventing juvenile delinquency. *Cochrane Database of Systematic Reviews*(4).

Ribisl, K. M., Walton, M. A., Mowbray, C. T., Luke, D. A., Davidson II, W. S., & Bootsmiller, B. J. (1996). Minimizing participant attrition in panel studies through the use of effective retention and tracking strategies: Review and recommendations. *Evaluation and Program Planning, 19*(1), 1-25.

Rosenbaum, P. R. (2010). *Observational Studies* (2nd Edition ed.). New York: Springer-Science+Business Media.

Schulz, K. F., Altman, D. G., & Moher, D. (2010). CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMC medicine, 8*(1), 18.

Yokum, D., Ravishankar, A., & Coppock, A. (2017). Evaluating the effects of police body-worn cameras. *Washington, DC: The Lab@ DC, 20*.