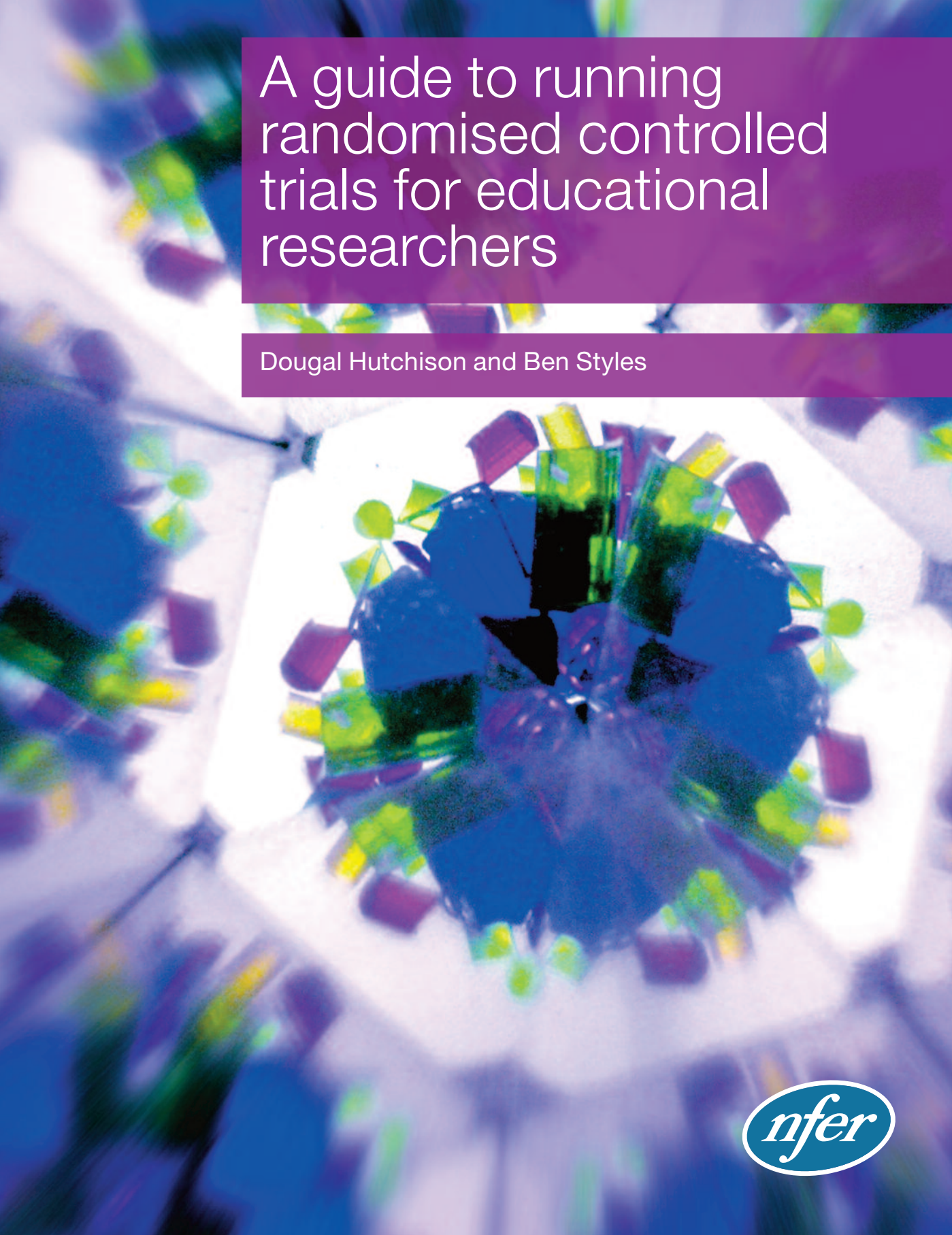
A guide to running randomised controlled trials for educational researchers

Dougal Hutchison and Ben Styles



nfer

About the authors

Ben Styles is a Senior Statistician at NFER. Having completed a DPhil in experimental science he learned his statistical theory through studying for the Royal Statistical Society exams and passing their Graduate Diploma. Over his 6 years at NFER he has worked on several randomised controlled trial evaluations including Rapid Reading (Harcourt), the Time to Read mentoring programme (Business in the Community) and, more recently, Rapid Writing (Pearson). He has been a regular proponent of RCTs in educational research both on the radio (BBC Radio 4's 'More or Less'), in a recent book chapter, through attendance at conferences, and to his long-suffering colleagues.

At the time of writing this Guide **Dr Dougal Hutchison** was Chief Statistician at NFER. During his 24 years there he has become an expert in all sorts of quantitative applications in education, including most recently, at the urging of his co-author, randomised control trials. He has a degree in mathematics from Edinburgh University and a Masters and PhD from the University of London. He has retired from NFER and is currently a visiting research fellow at the University of Oxford.

A guide to running randomised controlled trials for educational researchers

Dougal Hutchison and Ben Styles

National Foundation
for Educational Research

How to cite this publication:
Hutchison, D. and Styles, B. (2010). *A Guide to Running Randomised Controlled Trials for Educational Researchers*. Slough: NFER.

Published in September 2010 by the
National Foundation for Educational Research,
The Mere, Upton Park, Slough, Berkshire SL1 2DQ
www.nfer.ac.uk

ISBN 978 1 906792 68 8

Contents

Acknowledgements

We would like to thank the National Foundation for Educational Research for funding the production of this guide. We are indebted to Cath Haynes of Pearson Education for granting us permission to use data collected on her behalf in examples. Comments on the draft were received from Bette Chambers, Simon Gallacher and Michele Weatherburn.

1 The future is random – the 'why' and 'when' of randomised controlled trials

> *In one of the most tragic episodes in modern medical history, a seemingly minor change in the care of premature infants resulted in the blinding of about 10,000 babies between 1942 and 1954. The 12-year epidemic of retinopathy of prematurity was curbed only after a multicentre randomised trial demonstrated that the liberal use of supplemental oxygen had been the culprit. (Silverman, 1997).*

Although the penalty for failure and the reward for success are less extreme in educational research, the pattern of events in this example shows strong parallels with how educational interventions are evaluated. An exciting new educational policy is conceived and enthusiastically accepted. There is belated recognition that it might be necessary to show that it actually works. Evaluation then takes place long after the untested intervention is already in wide use. This approach shows a flagrant disregard for any kind of cost-benefit analysis of the intervention or the potential harm done to the individuals concerned.

For a researcher with control over an intervention, the above situation should not arise. The medical example starkly shows how rigorous evaluation put a stop to a detrimental treatment, but not necessarily that a randomised trial was needed. And yet randomised controlled trials (RCTs) are seen as the gold standard for evidence-based educational practice. In this introduction, I want to look in more depth at some of the reasons this is the case, and explain some of the statistical and epistemological power associated with RCTs. This is not to say that RCTs will be the best option in every case, however, and it is important to have a look at some circumstances where other approaches to research will offer advantages (Styles, 2009).

The evaluation problem and how randomisation solves it

For the purposes of this guide, we will accept the following conditions.

- There is a specific population of individuals at whom the intervention is aimed.
- Any improvements in the educational outcomes of the target individuals as a result of the intervention are measurable.
- The evaluation needs to discern the effect of the intervention on the outcome measure of interest.

Although questioned in academic debate, these conditions do represent a common scenario for researchers evaluating educational interventions.

The fundamental problem which the RCT attempts to solve is known as the evaluation problem. The ideal method of evaluation, were this possible, would be to compare outcomes for a set of individuals who has received the treatment with the outcomes for **the same set of individuals**, who had (somehow, magically) also not received the treatment. Obviously this would not be possible. Our next best alternative is to compare two different groups ensuring that they are equivalent in terms of any influential factors by randomising allocation of individuals to treatment or control between the groups. So, for example, if height is a key factor, we would expect that the mean height of the groups will be close enough that this factor cancels out. This logic applies to factors we know make a difference to outcomes and also those that we are unaware of. This strategy enables us to treat the groups as equivalent, provided the sample is large enough. It also enables a reasonable estimate of treatment effect to be determined.

Once we embark on selecting groups in any other way, the comparison is vulnerable to selection bias, meaning the bias that might come out of the way we have divided people up. Results have to be accompanied with a warning that selection bias may be present. It is difficult to say how significant this bias is since it may cover both known and unknown factors. Selection bias shows up as the difference in outcomes resulting

from the way in which groups were selected, rather than any effect of the intervention.

It can be helpful to view the evaluation problem and to explain the nature of bias using mathematical notation. This approach is adopted by Heckman *et al.* (1997). The measured outcome of an evaluation observed for an individual, Y, is defined as:

$$Y = DY_1 + (1 - D)Y_0.$$

Where D is the experimental condition and D = 1 if the person receives treatment and D = 0 if they do not. Y_1 and Y_0 are the outcomes in conditions 1 and 0 respectively. The gain from participating in the programme is $\varnothing = Y_1 - Y_0$. The evaluation problem can be restated as the problem that we cannot simultaneously observe Y_1 and Y_0 for the same person.

To work out the effect of the intervention, in a simple formulation, we can subtract the mean outcome among non-participants from the mean outcome of participants (Heckman and Smith, 1995). This leads to the equation (where E means 'the expected value of', or mean; and | means 'given'):

$$E(\varnothing) = E(Y_1|d = 1) - E(Y_0|d = 0)$$

However, this way of expressing the problem does not take selection bias into account. Selection bias arises when one group differs from the the other in some important way: for example, the treatment could be given to those that appeared to need it most. The researcher could attempt to cut down on this by matching the groups on characteristics thought to be important. However, matching on factors you can measure leaves open the possibility that differences still remain on factors you have not measured. If people are put into treatment and non-treatment groups using a process that is not random, even one that relies on equal distribution of the visible factors that are suspected to be important for the experiment, then selection bias will exist. Of even greater concern is the fact that we will not know how much of it there is, and hence the simple subtraction of mean non-participant outcome from mean participant outcome will not provide a causal conclusion.

In any experiment, some people agree to be randomly allocated to either receiving or not receiving the treatment, while others are not prepared to be randomly allocated to a group. It is easy to see why somebody suffering a serious illness might wish to be certain that they would receive an experimental treatment rather than run the 50 per cent risk of not receiving it. However, the treatment will often not be available outside the trial. Since we do not know whether the treatment is helpful, harmful or makes no difference, denial of the treatment through randomisation is an ethically neutral scenario.

The group of people who agree to random allocation are designated D*, and they are divided between those who get the treatment (R=1) and those who do not (R=0). If the treatment to be tested is already widely available, it may be considered ethically unacceptable to deny treatment to people who refuse random allocation. A group of people who get the treatment but refused random allocation cannot be taken into account when measuring the outcomes.

Randomisation of the D* participants removes the problem of selection bias because random allocation to the R=1 group or the R=0 group should produce a situation where the two groups have very similar average 'amounts' of all the factors that could affect the outcomes. It is important to note that this applies both to visible factors, including those that are likely to be relevant to the experiment, and invisible factors as well, which could also be relevant.

Using mathematical notation, the estimate of the causal effect can be summarised:

$$E(Y_1 - Y_0|d^* = 1) = E(Y_1|r = 1 \text{ and } d^* = 1) - E(Y_0|r = 0 \text{ and } d^* = 1)$$

However, there is no free lunch in statistics. While this approach, the heart of randomised trials, does deal effectively with selection bias, it creates a new bias, known as randomisation bias. This occurs because only some people agree to be part of the randomised D* group, and this agreement is not random. That is, there could be certain shared characteristics of the people who decline to take part, and this cannot be known. So randomisation is really randomisation within a self-selecting group. Randomisation bias and other methodological

problems associated with RCTs in educational research are considered by Styles (2009).

Common objections to RCTs

The passing of the US Education Sciences Reform Act of 2002 provoked a renaissance of scientifically-based educational research in the United States. Resistance to the use of RCTs in the UK is discussed by Oakley (2006), and objections to the use of randomised trials in educational research are well debated (Cook, 2002; Cook and Payne, 2002). Table 1.1 suggests responses to some of the common objections to using RCTs in educational research.

Table 1.1 Responses to objections raised

Objection	Response
Randomisation is unethical.	If we do not know whether the intervention works or is even detrimental to outcomes, we are in an ethically neutral scenario when randomising. If we know it works (for example, through the use of previous randomised trials and a meta-analysis) then no evaluation is necessary, and we should not be randomising. In educational research, we rarely know something works in advance. Contrast randomisation with the cost (both monetary and educationally) of rolling out an untested intervention.
Participants will think randomisation is unethical.	This is a genuine problem and can reduce the generalisability of results if those agreeing to be randomised do not form a representative sample. The set-up of a trial should incorporate discussions with those who will be running it regarding the principles behind randomisation. If participants understand why they are being randomised they are more likely to take part.
Limited generalisability.	The critics of the generalisability of RCT methods need to ask themselves: is it better to have the 'right' answer among a relatively narrow group of participants or the 'wrong' answer amongst everyone (Torgerson and Torgerson, 2008)?

Table 1.1 Responses to objections raised (continued)

Objection	Response
We cannot practice blinding (see Chapter 2) in educational research trials.	In a drug trial, patients, doctors and researchers can be made unaware of what treatment patients are receiving (a double-blind placebo trial). In educational research, this is rarely possible since the intervention is clearly visible to all concerned. We are, however, often able to blind those involved in the measurement of outcomes. The lack of blinding is a problem for educational research trials due to the propensity for intervention group participants to 'try harder' (known as the Hawthorne Effect). However, other evaluation methods will rarely improve on this since they will also not operate blinding.
The interaction between pupils and their education is too complex for a simple experiment.	This can actually be considered as an argument for running an RCT. Since the unknown factors that are balanced out through randomisation are so prevalent in educational research, any non-experimental methodology will be subject to selection bias.
It does not tell us how the intervention works.	True. It tells us whether an intervention has worked. This is a good argument for qualitative work happening in parallel with any trial that also seeks to clarify how the intervention is causing any measured effect. If we are still in the dark, does it really matter if we know it works?
Standard intervention implementation is assumed.	The randomised trial is useful in assessing the effectiveness of an intervention as if it were used for real. If there is poor fidelity of implementation in the trial, the same is likely to happen if used for real. The results, therefore, necessarily reflect how the programme was implemented. In fact, the intervention could be administered with greater fidelity than for real due to the Hawthorne Effect. This should be borne in mind when interpreting results.
The complex interplay of systems and structures within education means that large-scale policy changes are difficult to trial with an RCT.	Take the introduction of a new examination system, for example. This requires the complex interplay of pupils, teachers, schools, markers, funding streams and examination boards. It is unlikely, for practical reasons, that the intervention could be rarefied sufficiently for a cluster-randomised trial at school level. Instead, some kind of observational study (Styles, 2009) could be used but would not demonstrate causality. An attempt should be made, even if simply conceptual, to harness the intervention into a form that it is possible to trial. If a randomised trial is impractical, then there is a genuine case for an observational study.

When to use an RCT

An RCT should be considered as the first choice to establish whether an intervention works.

- It eliminates selection bias and can generate a causal conclusion.
- It avoids potentially misleading results from non-experimental work which has inadequately controlled for selection bias.
- It provides a quick and digestible conclusion of programme effectiveness that avoids lengthy caveats.
- Its results can be incorporated into future meta-analysis.

One situation where an RCT seems genuinely inappropriate is the evaluation of policy changes that affect many parts of the educational system (see Table 1.1). Otherwise, random assignment is the best mechanism for justifying causal conclusions.

A huge problem for RCTs at a national level is the regular rolling out to schools of initiatives which have not been properly trialled. Evaluation is either a late afterthought or not considered important enough to delay roll-out. Schemes that have suffered this fate include the School Fruit and Vegetable Scheme and the National Literacy Strategy. The exposure of children to educational harm when initiatives are not properly tested is a very real risk (Oates, 2007). Researchers in control of their small-scale interventions are sufficiently insulated from the requirements of politicians, and the demand for quick roll-outs of interventions, not to have to cut corners when evaluating.

Reporting the results of an RCT

Randomised trial reporting should follow the Consort Statement at http://www.consort-statement.org/.

2 Ed versus med: differences and similarities between educational and medical experiments

It is a truism to say that the current configuration of any development is heavily conditioned by its origins and history. Thus, for example, the current shape of the motorcycle is determined to a large extent by its origins as a pedal bicycle with an engine attached and, supposedly, the standard railway gauge comes from the wheelbase of Roman chariots. In a comparable fashion, many of the features of how RCTs are used in education and social science are conditioned by the their origins in medical experimentation. This will be illustrated by three examples of RCTs applied in medicine and social science.

In this chapter, we discuss and compare some of the aims and practicalities of experiments in the two areas, medicine and education. One of the aims of this discussion is to argue that conventions, both practical and ethical, need not be exactly the same in an educational setting as in a medical experiment.

Example 1: medical experiment on a new treatment for a disease

A pharmaceutical company has produced a new drug, and small-scale prior experiments suggest that it will represent an improvement on the existing drugs for treating a serious and widespread disease. The company propose an experimental trial, involving randomising between patients in an experimental group: those who receive the new drug, and a control group, who get the old one.

There are standard legal guidelines for such trials, so the company asks an ethics committee to scrutinise the details of the proposed experiment. The ethics committee assesses the details of the trial, and requests a few alterations to improve the study. These are taken on board and the trial can start.

Patients presenting with the disease in question are asked if they wish to take part in a randomised experiment and, if they accept, they are randomly allocated to receive either the new drug or the old one. This randomisation procedure is carried out by a specialist outside body to ensure those involved in the experiment do not influence the randomisation procedure. The patients are not told whether they are in the 'control' or 'experimental' group, and neither are their carers. The appearance of the pills in which the drug is taken is similar, so it cannot be used to deduce which group the patients are in. Measurements of symptoms are made at the beginning and end of the experiment: again, these are made by staff who do not know which group the patients are in. This process of disguising the group membership is referred to as blinding.

If the trial gives results that show the new drug is superior to other chemicals already available on the market, then the company will consider introducing it to replace or supplement the previous treatment.

Example 2: teacher's study of learning skills and pupil progress

A blend of reading the background literature and personal experience convinces a teacher that the pupils in her school would perform better if they were given formal tuition in study skills. Realistically enough, she considers it will not be possible to teach one half of the class in one way, and the other half in another. She is convincing enough to persuade her fellow teachers to take part in a study, and it is agreed that three classes will be involved as an experimental group, and three as a control group.

She considers this is sufficiently within her normal realm of decision making as not to require any kind of formal consent from the pupils or their parents. Standard internal examinations take place at the end of every term, so the two sets of examinations, a term apart, are taken as the before and after measures. If her experiment proves successful, in that she considers the pupils' performance has improved as result of the new technique, she may adopt it for use in her own class. She may also recommend its use to other teachers.

Example 3: government initiative on teaching reading

Some medium-scale academic studies suggest a new method of teaching reading is superior to the old one. This new method hits the headlines and is eagerly seized upon by the media and politicians. However, there are a number of potentially quite serious objections raised by the academic research community. The government department responsible for education decides it would be a good plan to carry out a study to look into these. Research organisations bid for this work and a team is appointed. A sample of schools is drawn. The study team approaches them, explains the study, and asks if they would like to take part. Those that agree are then randomised into two groups, one being expected to administer the new technique, and the other, the old technique. It is not feasible to give a treatment to only some pupils in a school, and potential contamination (see below) is large, so the selected schools are some distance apart and all pupils in the school or year are taught in the appropriate way. If the trial proves 'successful', it is likely that the new method will be introduced more widely in schools.

Compare and contrast

These scenarios are hypothetical, though representative of many studies of their type. A closer look at these relatively simple examples shows a number of important features and contrasts.

Population involved

In the first example, the individuals involved are a relatively small proportion of the entire population, and a selected group having a presenting problem, that is, a disease. In the second, the research population is simply the school. In the third, the aim is to have the studied schools as representative of the whole population as possible.

The medical development is attempting to deal with some kind of serious problem and it is expected that the treatment will make a substantial impact, for example, be a cure for malaria or tuberculosis. It is also possible that the treatment could have a substantial negative impact, for example, some of the first batches of polio vaccine actually infected those inoculated with the disease. For teachers educating

children, results are likely to be relatively smaller incremental changes in performance, attitude or behaviour.

Ethics

Given the possibility of a medical intervention having major impacts, it is particularly important to 'get it right'. If a treatment actually makes an impact different from existing practice, then someone among those involved is going to be disadvantaged. If a new treatment is shown to be very effective, then the control group is losing out by not experiencing it. Conversely, if the treatment is a disaster, the experimental group is disadvantaged. Also, if it is successful, the sooner it can be available more widely the better. It is important that all concerned are aware of the possible potential benefits and problems, and for this reason the role of the ethics committee is central.

The necessity of an ethics committee in small-scale classroom research, as outlined above, is less evident. Even in these days of prescriptive national assessment, and allegations of teaching to the test, a teacher has a substantial degree of freedom in determining what goes on in the classroom. It is generally accepted that teaching style is a largely personal quality, and the teacher, in this example, may feel that a change with the impact of teaching learning skills is something that she could introduce on her own initiative without wider consultation. In this circumstance, the teacher might well argue that if she could do this anyway on her own initiative she should not have to consult an ethics committee simply because she is carrying out research.

This is a somewhat controversial aspect. Many experts would maintain that an ethics group is a *sine qua non* for any type of human experimentation, and it may be that this becomes necessary in the future. It will certainly be advisable to do this when planning to publish results or as part of a degree project. However, it is at least clear that the impetus for this comes from the technique's roots in medical experimentation.

Risk

This consideration is closely allied to the previous one, but sufficiently different to merit separate consideration. There is likely to be an element of risk in an experimental trial. Either the control group could be denied

an effective treatment, or the experimental group could be exposed to an untried or risky procedure. Further, if the treatment is in fact effective, those not in the trial may not even have a chance of receiving it while the trial is going on. So there is a tension here.

On the one hand, the study needs to minimise the risks that the subjects are exposed to. This argues for a short study, involving a relatively small number of people. On the other hand, the study has to be able to ensure that it is possible to reach a conclusion with a sufficient degree of certainty. If the study is too small to reach a conclusion, then arguably all the contributions will be essentially wasted. This argues for a large-scale study.

To balance these competing demands most efficiently it will be essential that statisticians are fully involved in the design, both in the research team and in the ethics committee. Similarly, a large-scale education trial must be designed with a sample size sufficient to detect an important effect if one is there. In some situations a quite small effect may still be worth pursuing, so quite a large sample would be required.

The considerations of the teacher researcher are rather different. Her sample size is more or less fixed. However, it will still be necessary to involve statistical advice to ensure that the proposed design will be strong enough to be able to detect the size of effect that would be of interest. If not, it may be necessary to expand the study to include more than one school, for example.

Randomisation

If those involved in the medical experiment are also involved in selecting group membership it is possible that the two groups are different in potentially important aspects. Subjects who are more voluble, or just more desperate, may be more likely to get themselves into a treatment group. Experimenters may be more likely to select those who are less seriously affected, believing them more likely to show a favourable outcome, while administrators may be more likely to select the worst cases, on the basis that they need it more. For this reason subjects are assigned using a random process. To ensure that there is no influence, even a subconscious one, on selection, one approach is to have the

randomisation carried out secretly by a third party. This approach is also recommended for small-scale teacher research. Large-scale educational research will typically sample schools from a standard list of schools, and then pupils within them according to a published and standardised sampling scheme in which there is no role for researcher preferences.

Expectations of experimenter and subjects, and contamination

These can influence outcomes. It is generally considered that the very fact of appearing to experience a treatment can be likely to induce improvement (the placebo effect). Something similar can happen in social research (the Hawthorne Effect). For this reason, the medical trial was conducted 'blind': neither patient nor experimenter knew which group each subject was in, nor were they able to influence the assignment to control. In contrast, in the small education example, while it might be possible to disguise from the pupils which group they were in by giving them an alternative treatment, it is obvious to the teachers which group their classes and their pupils are in. Realistically, it is going to be difficult to hide this information from pupils too.

If pupils know which group they are in, there is likely to be 'contamination' under which the effect of the treatment is diluted by being made available to the control group. This type of effect is less likely to happen in the medical example. Under normal circumstances, a teacher is unlikely to be able to teach one part of her class in a different way from another part (though this may be possible using some kind of software in a computer room). For this reason allocation is likely to be at a whole-class level: all pupils in one class have the same treatment, and all pupils in another class share the same treatment, though this is different from that in the other class. (There could be some contamination when pupils compare notes in the playground, or the 'control' teachers adopt some aspects of the research strategy.) This, in turn, means that the design is much weaker, and it is unlikely that any but the strongest effects can be detected with confidence.

In large-scale education studies contamination can be minimised by introducing initiatives at a whole-school level, and also by making sure that schools taking part are not closely adjacent.

Intention to Treat

Being human, some patients are likely to drop out during the course of the study. Since this process is also likely to take place in real life, the experimenters introduce a concept known as Intention to Treat, so that the outcome is studied for each group no matter whether they complete the experiment or not. This is less likely to take place in the small-scale education experiment described, but there is still the possibility that some pupils may leave the school or be absent at the time of testing. In larger-scale education experiments pupil or school dropout or non-cooperation will occur. There are techniques to make adjustments for such non-response, but it is best to always plan the study and contact with schools to minimise any such dropout first: non-response adjustment techniques may then be used.

In the small-scale education example, the teacher was able to use the pre-existing structure of termly tests as pre- and post-tests. This is very convenient, and it is well worth piggy-backing on well-established structures, if possible. By contrast, in the medical example, if it is not standard practice to test at all stages, it may not be worth doing a pre-test.

Inference

In the medical study, if it is carried out properly, the experimenter will be able to make strong claims about internal validity, and, depending on circumstances, about external validity. By contrast, the teacher may be able to draw conclusions about internal validity, for example, of her own teaching practice, but not be able to make much in the way of external validity claims.

One point that it is important to remember is that any innovation is likely to be a zero-sum game: if a teacher is teaching extra mathematics, then she is probably teaching less of something else. Or if it's done in her spare time it may be at the expense of her energy or pupils' morale.

In summary, it may not be unfair to say that the medical experiment has a lot of weight in terms of people's health and well-being and indeed sometimes lives, not to mention financial investment, so the aim of the sample design must be to take account of all circumstances, and to make the best possible decision.

By contrast, while education has a very important long-term aim, the impact of individual changes is likely to be relatively small. The teacher's aim, rather than getting the best possible, has to be getting the best she can. In a large-scale education sample, in addition to ensuring that it is large and well enough designed to be internally valid, it will be important to sample in a way that is sufficiently representative of the population as a whole so that the results will also be externally valid.

Finances

Is the game worth the candle? In every type of study, there must be an underlying appreciation of the balance between improvement and cost. Thus, it will not be sufficient to show simply that a new drug is better than the old one: it will be necessary to show that it is sufficiently better to justify the substantial costs in changing the manufacturing process, advertising, training staff to use it, and so on. Similarly, a teacher will have to decide whether the new practice is more onerous than the old one and, if so, whether the improvement justifies the increased expenditure of effort. Finally, a government research initiative may well wish to conduct a formal cost-benefit analysis to see whether observed improvements justify the increased expenditure. For all three approaches, it is also likely to be important to compare this with possible alternatives in terms of value for money.

3 The simplest design

The very simplest design of an RCT involves:

- predicting the sample size required
- randomising the individuals involved into two groups
- applying the intervention to one of the groups
- measuring the outcome for the individuals involved
- comparing the results and interpreting the differences.

These stages are considered in turn in this chapter.

Predicting the sample size required

We are usually concerned with results that can be generalised to a wider population, for example, a particular year group within a school or all schools of a particular type. For a trial where the individuals concerned can be a simple random sample of the population, sample size calculation is relatively straightforward. This situation might arise, for example, if we were sampling from a list of pupils in a particular year within a school or a list of teachers in England. Here the total sample size for the trial can be approximated using the formula (Lehr, 1992):

$$\frac{32}{\text{Effect size}^2}$$

The effect size is a measure of the effect of the intervention in standard deviation units of the outcome. To predict sample size, we, therefore, need to estimate what size of effect we would need to measure to be convinced that an intervention has worked. This should ideally be built on prior knowledge of the outcome measure and how large a difference would be educationally relevant. In practice, rules of thumb are used and it is often the case in educational research trials that trials are designed to detect effect sizes as small as 0.2 (that is, total sample size of 800; 400 in each group). However, if we had a great deal of confidence in the effect of the intervention, it would often still be justified in designing a trial to

detect an effect size of 0.5 (that is, total sample size of 128; 64 in each group).

An educational research trial often requires asking consent. It is important to randomise the individuals that have agreed to take part rather than those who have been asked to take part. For a trial on pupils, the consent of school headteachers is often sufficient. If a trial is run on a new teaching method for pupils within a year group at a school, say, this may happen during the course of normal teaching and consent may not be an issue. However, requirements for consent are increasing, and it would be advisable to seek guidance, especially if the data is to be used for any kind of formal research purposes. For larger trials where consent is an issue, please refer to Chapter 5.

In educational research, we are often concerned with sampling clusters of individuals, for example, in schools or classes, and sample size considerations become more complex. Chapter 6 addresses these scenarios.

Randomising the individuals involved into two groups

It is important that you make every effort to ensure that your experiment is genuinely random, and you have to watch out for unconscious biases. These can come in at any time during the trial but, at this stage, we are dealing with the randomisation process. If you have a pet theory, or if you feel that the results are in some way reflecting on your own performance, it can be very easy to let what you would like to happen influence the allocation, for example, avoiding an un-cooperative pupil or redoing the allocation if you do not like the way it looks. This may seem easy, but renders your experiment worthless as a piece of scientific research.

To avoid this happening, the best approach is to plan out your study, write down in advance the plan of action and make sure you follow it to the letter. Get a mentor to hold the plan and sign off all the stages as you proceed. This way, not only will you ensure that you do not influence the sampling subconsciously, you will also be able to rebut any suggestions that you did.

How should you carry out the randomising? We shall give a description of how it can be done for two groups. The process for more than two is similar. It can be carried out in Excel or SPSS. The advantage of using SPSS syntax is that you end up with an audit trail of how the randomisation was carried out. It can also be done by hand using a random number table to generate a random number for each individual.

The randomising process

Step 1

Produce a list of the names of the individuals who will be the experimental subjects. This can be in any order and alphabetical is fine. Number these in sequence 1, 2, … . The data set should look like this:

Alice Bacon	1
Cedric Duncan	2
(etc.)	3
	4
	18
	19
	20
William Younger	21
Xavier Zylstrom	22

Copy and paste or write these directly into Excel or SPSS. In SPSS name the variables NAME and ID.

Step 2

Create a set of uniform random numbers, one for each of the subjects. In Excel use the formula =RAND() in the cell to the right of the pupil number and copy and paste this down to the bottom of the list. Or use the following SPSS syntax:

```
compute vrand = RV.UNIFORM(0,1).

execute.
```

The resulting data file should look like this:

Alice Bacon	1	0.065737
Cedric Duncan	2	0.473664
(etc.)	3	0.817567
	4	0.534988
	5	0.333037
	6	0.970875
	7	0.17347
	8	0.863593
	9	0.559949
	10	0.468753
	11	0.838642

	12	0.515478
	13	0.723583
	14	0.516084
	15	0.675724
	16	0.495068
	17	0.271099
	18	0.674255
	19	0.065279
	20	0.312911
William Younger	21	0.31516
Xavier Zylstrom	21	0.395487

Step 3

Sort the cases by the random number variable. In Excel use the sort cases dialogue box or use the following SPSS syntax:

```
sort cases by vrand.
```

Step 4

If there is an odd number of individuals, decide which group will have one extra before allocation. Allocate the first half to the intervention group and the second half to the control group. In Excel this can be done by writing 1 in a new column for the first half and 2 for the second half on the list. If the file is then required in the original order, cases can be sorted back using the pupil number. In SPSS syntax:

```
if $casenum le 11 group=1.

if $casenum ge 12 group=2.

sort cases by id.
```

The resulting data file should look something like this:

Alice Bacon	1	0.065737	1
Cedric Duncan	2	0.473664	1
(etc)	3	0.817567	2
	4	0.534988	2
	19	0.065279	1
	20	0.312911	1
William Younger	21	0.31516	1
Xavier Zylstrom	22	0.395487	1

Applying the intervention to one of the groups

The pupils are then allocated according to the appropriate group, and the intervention carried out with those in group 1. The remaining pupils make up the control group who, in the simplest design, carry on as they would have done anyway.

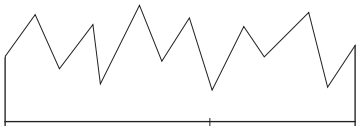
Measuring the outcome for the individuals involved

One crucial aspect of the outcome being measured is that it does not relate directly to the intervention being used. For example, if new reading materials contained examples from football and ballet, the final reading test should not be designed around questions about football and ballet. In this scenario, any improvement seen in reading performance could be attributed to subject matter rather than general reading ability. Also, if a researcher or intervention developer designs the test, they may unintentionally create a test that favours the experimental group. This problem is often solved by using a pre-developed commercially available test. This has the added advantage that you do not have to develop the test yourself.

Comparing the results and interpreting the differences

Once you have conducted the experiment, and tested the participants, the results are entered into your data set. It can be easy to make mistakes at this stage, so it is a good plan to do this twice. In more formal research this is done using a process known as punching and verifying or, more recently, scanning. Smaller-scale research can do this by entering the data twice, and comparing the results. The resulting data could look like this.

Group	rscore88
1	36
1	-1
2	12
2	43
1	30
2	-1
1	18
2	42
2	31
1	29
2	12



1	27
2	-1
2	34
2	14
1	19
1	12
2	25
1	-1
1	-1
1	35
1	-1

Note that the second case has apparently a score of -1. This is because 'missing' values have been coded as -1. A missing case arises where it has not been possible to record a score for an individual who took part in the project. Other cases also have missing scores.

Ideally, we should ensure that there is no missing data in the first place. In particular, a supplementary testing session can be arranged to pick up those off sick the first time. Even after this, there may be some for whom there is still no score. How we deal with the missing cases depends on how we interpret the fact that the individual is 'missing'. It may be that we can assume that those not tested are no different from the rest. For example, they could just be off sick on that one day. In this case, it may be best just to exclude them from the analysis. This can be done in SPSS by the command:

```
missing values rscore88 (-1).
```

This command tells SPSS to ignore any cases with this value when dealing with the variable RSCORE88.

Alternatively, no result recorded may be a very important finding. For examinations, it could indicate that the individuals were truanting, and this may be happening at different rates in intervention and control groups. Or, in a project aimed at changing attitudes, it could indicate that the project had been so unsuccessful that individuals had dropped out of the intervention group.

In these situations, simply excluding individuals from the analysis could bias the result. There are sophisticated methods for dealing with missing values when other background data is available (Shadish *et al.*, 2002). If there is suspicion that individuals are missing for a reason that might in some way impact on the results then the analysis may be compromised without (or even with) these methods.

A common practice is to include missing individuals with the mean score of the group. This should not be done since it probably misrepresents what they would have scored and gives us the impression of greater precision than was actually available. Chapter 7 has a more detailed consideration of missing data.

In this example, we shall exclude the missing cases, assuming they are missing for a reason that is completely unconnected with the experiment and the outcome measured. Now we shall conduct a t-test to determine whether there is a difference in outcomes between the two groups. The following SPSS syntax can be used for this:

```
missing values rscore88 (-1).

T-TEST GROUPS = group(1 2)

/VARIABLES = rscore88.
```

The output from this is shown in Tables 3.1 and 3.2.

Table 3.1 Group statistics

Group	N	Mean	Std Deviation	Std Error Mean
rscore88 1	8	24.88	10.357	3.662
rscore88 2	8	26.63	12.917	4.567

Table 3.2 Independent samples test

	Levene's Test for Equality of Variances		t-test for Equality of Means						
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std Error Difference	95% Confidence Interval of the Difference	
								Lower	Upper
rscore88 Equal variances assumed	.940	.349	-.299	14	.769	-1.750	5.853	-14.304	10.804
Equal variances not assumed			-.299	13.369	.770	-1.750	5.853	-14.360	10.860

The first point to note is that, because of missing data, the numbers in each group are down, from 11 to 8. This is likely to reduce the effectiveness of the experiment.

In this chapter we shall use the term 'significance' (properly 'statistical significance', and sometimes abbreviated to 'sig.'). This answers the question: 'Suppose there is actually no difference. What is the probability of an apparent difference of this size arising by chance?' This probability varies between 0, meaning that it is impossible that this could have happened by chance alone, and 1, meaning that it was certain that this could happen. By convention, probabilities below 0.05 are treated as **statistically significant**. This means that there is only 1 chance in 20 that the difference is due to chance.

There are two types of t-test in the situation, depending upon whether you treat the variances of the two groups as equal or unequal. The first part of Table 3.2 deals with whether they should be treated as equal or unequal. Sig. has a value of 0.349. This is substantially larger than 0.05 and means that it is quite likely that the difference in variance seen could have happened by chance, even where there is actually no difference. This means that, on the basis of the data we have here, there is no reason to believe that the variances of the groups are not equal. For this reason, we can treat the variances of the two groups as equal as far as the t-test is concerned.

The next aspect, and the one which is of most interest, is to assess whether there is actually a difference between the two groups in average performance. To do this we set up a **null hypothesis**, namely that there is really no difference between the two groups, and see what the possibility is that this observed difference could have arisen by chance given this assumption. There are two parts of Table 3.2 that answer this question. The first is Sig.=0.769. This is much larger than 0.05 and implies that it is likely that the observed difference could have occurred by chance and we can reasonably accept the null hypothesis as true. Alternatively, we can look at the extreme right of Table 3.2, the part labelled '95 per cent confidence interval of the difference'. To say that the 95 per cent confidence interval extends from -14.3 to 10.8 may be interpreted to say that there is only a 5 per cent (= 100 – 95) chance that the actual value lies outside this range. The important question is whether the estimated

confidence interval contains the value 0. If it does not, then we assume that this study provides us with evidence that the difference is not zero. If the confidence interval does contain zero, then this study does not provide us with any evidence to abandon our null hypothesis. In this case, the confidence interval contains 0 and we, therefore, decide that the experiment does not provide evidence of any effect from the intervention.

Real-world example of the simplest design

An education body wishes to upgrade their registration database in order to achieve a comprehensive coverage of registrants' ethnicity and disability data. In order to assess the feasibility of this, a pilot study was conducted where this data was requested from a sample of respondents. One possibility of particular interest was whether including a pre-paid envelope might improve the return.

The proposed research question lends itself beautifully to randomisation. The population was defined and a single stratified random sample of 6800 teachers was drawn from the database. The sampled teachers were then randomly allocated to one of two groups. Each group was mailed a letter requesting ethnicity and disability information. In addition, one of the groups was sent a pre-paid reply envelope:

- group 1: data request form and cover letter, and pre-paid envelope
- group 2: data request form and cover letter, unpaid envelope.

The proportions replying were pretty low, but they appeared to be affected substantially by whether or not a pre-paid envelope was included. The analysis output is shown in Tables 3.3 and 3.4. The variable RESPOND is coded (0 = No; 1 = Yes).

Table 3.3 Group statistics

	Group	N	Mean	Std Deviation	Std Error Mean
respond	Not prepaid	3400	.2579	.43757	.00750
	Prepaid	3400	.3344	.47185	.00809

Table 3.4 Independent samples test

	t-test for Equality of Means						
	t	df	Sig. (2-tailed)	Mean Difference	Std Error Difference	95% Confidence Interval of the Difference	
						Lower	Upper
Equal variances assumed	-6.929	6798	.000	-.07647	.01104	-.09810	-.05484

Among those receiving the pre-paid envelope, approximately a third, 33.4 per cent, responded. Of those that did not receive the pre-paid envelope, just over a quarter, 25.8 per cent, responded. Table 3.4 shows sig.=.000 hence this difference is statistically significant at the 0.1 per cent level, and it also seems to be an effect that is practically useful. It is, therefore, clear that including a pre-paid envelope helps response.

4 Testing before the intervention

So far we have looked at an 'after only' design for experiments:

- set up the two groups for the experiment
- carry out your experiment
- measure whatever it is at the end.

There is a more powerful design, still under the randomised trials umbrella, which can be schematised as:

- set up the two groups for the experiment
- measure whatever it is at the beginning
- carry out your experiment
- measure whatever it is at the end.

This is a more powerful design than the simple 'after only' design, and often a substantially more powerful one. It is more powerful because the use of baseline data in the analysis can explain some of the variation in follow-up outcomes. Any genuine difference in progress between the groups of the trial is then less masked. Baseline data also has a role in correcting for any imbalance in outcomes between the groups of the trial that might have resulted at randomisation. Such imbalance could have arrived by chance even if the randomisation were carried out properly or through some fault in the randomisation process. In this latter case, the trial cannot truly be said to be randomised.

The use of baseline data in the design and analysis of a randomised trial will be considered using a real-world example. However, first, it is important to visit the issue of sample size. Since the use of baseline data can make our analysis more sensitive, it is possible that it can reduce the sample size required for the experiment. This issue is considered in Chapter 6.

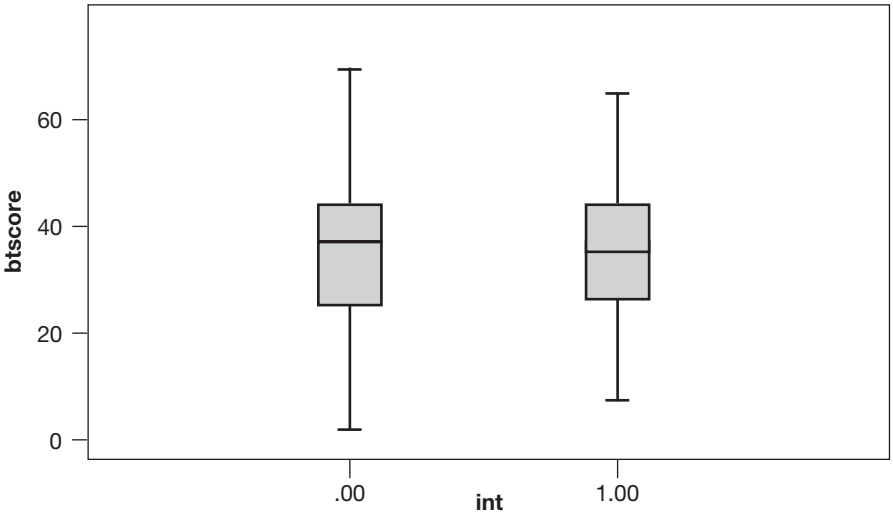
Example of analysis with baseline data

This example uses data from a randomised controlled trial of reading materials for struggling readers. A box-and-whisker plot of the post-intervention measure obtainable using the following syntax in SPSS is shown in Figure 4.1. The variable BTSCORE is the outcome and INT has a value of 0 for the control group and 1 for the intervention group:

```
examine vars=btscore by int/

plot=boxplot.
```

Figure 4.1 A box-and-whisker plot of the post-intervention measure for intervention and control groups



It can be seen from the plot that there does not seem to be much difference in outcome scores between the two groups: the median and inter-quartile ranges are very similar. Table 4.1 shows the output after analysing the post-intervention measure using a t-test.

Table 4.1 Independent samples test

	Levene's Test for Equality of Variances		t-test for Equality of Means						
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std Error Difference	95% Confidence Interval of the Difference	
								Lower	Upper
Equal variances assumed	3.648	.056	.888	784	.375	.798	.899	-.966	2.563
Equal variances not assumed			.882	742.101	.378	.798	.905	-.979	2.576

This confirms our suspicions that there is not much going on. The significance value of p=0.375 suggests that we should accept the null hypothesis that there is no difference between the groups.

However, we have more data at hand. We are able to explain some of the post-intervention outcome variability by using pre-test data in our analysis. In other words, since we know which pupils are better or worse readers on the basis of their pre-test score, we can do a more sensitive test of whether the intervention has worked. The pre-test was a different test to the post-test that tested the same reading construct. Using a different test has the advantage that pupils do not learn directly from the pre-test for the post-test. This analysis is done using a regression model:

```
regression vars=btscore atscore int/

 dependent=btscore/

 method=enter.
```

Table 4.2 shows part of the outcome from this analysis.

Table 4.2 Coefficients

Model		Understandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std Error	Beta		
1	(Constant)	10.647	.616		17.278	.000
	atscore	.780	.016	.872	49.066	.000
	int	.691	.443	.028	1.559	.119

Here, the post-intervention outcome BTSCORE is regressed on ATSCORE, the baseline measure, and INT, the grouping variable. We can see from the large value of t and the $p<0.001$ significance value for ATSCORE that this is a highly significant predictor of BTSCORE, as we might expect. However, even with this more powerful analysis, the significance value for INT is $p=0.119$, that is, it is not statistically significant at the 0.05 level and we can accept the null hypothesis. The coefficient of INT (0.691 score points) tells us the size of the difference between the two groups, after taking into account any differences in ATSCORE.

From this example, we can begin to see that the use of baseline data has helped. The value of t has increased from 0.89 in the t-test to 1.56 in the regression analysis. Perhaps there is something very slight going on that we might have detected with a bigger sample? The important point to note here is, even if there is an effect, it is very slight in terms of size (effect size=0.691/12.574=0.05) and probably not of educational significance (12.574 is the BTSCORE standard deviation).

5 Clustered data

So far we have talked in terms of allocating individuals (pupils and patients) at random to groups. However, this can be more of a problem than you might at first anticipate. Suppose you want to compare two methods of teaching a topic, or two reading schemes, for example. Can you allocate the pupils randomly within a classroom? Well, yes, of course you can – but is it a sensible procedure? What would be the problem?

If you are trying to teach half the class using one method, and the other half by another method, while both are sitting in the same room, then it is probable that, in most cases, one half will hear what the other is being told, and vice versa. This is called **contamination**. Perhaps this might not happen if you are using some kind of individualised instruction, using either workbooks or computer-aided instruction. However, even then they may overhear some of what is happening with their neighbour.

In such a situation, it is usual to allocate interventions to complete classes. This could certainly be expected to cut down on the amount of contamination, and you could expect to reduce this still further by using different schools. However, this immediately gives rise to other problems. Suppose you divide the pupils into two groups A and B, and ask teacher A to deliver intervention A and teacher B to deliver intervention B. We then find that pupils in group B (say) do substantially better than those in group A. Can we assume from this that intervention B is better than intervention A? Not necessarily: it could be that teacher B is more effective on this topic than teacher A.

This raises the question of the **unit of allocation**. If we have 200 pupils, 20 to a class, then we have 10 classes. If we then randomly allocate classes to interventions, we have only 10 separate units of allocation. In this case, the unit of analysis should also be the unit of allocation. It is unlikely that we would have sufficient classes to be able to make any conclusions from this kind of experiment. For this reason it is virtually impossible to run a cluster randomised trial within a single school.

Should we take the whole class or just a sample if allocating interventions at classroom level? The basic statistician's mantra is 'other things

being equal, the more the better'. However, 'other things being equal' is quite an important qualification and the improvement in accuracy from additional data collection may not be worth the cost of collecting it. In this case, it depends crucially on the administrative procedures used in the study. If we are envisaging pupils being withdrawn singly or in small groups for additional attention or computer experience, then the fewer withdrawals the better. However, if you are planning to give, for example, a written test to a group of children, it is probably going to be less disruptive to the teachers and pupils involved to test the whole class. In this situation, unless buying or printing the actual tests is expensive, you have effectively got 'free data' if you test the entire class!

So, in carrying out an RCT comparing subjects already organised into clusters (a **cluster randomised trial**), the steps are essentially as before for the individual-level allocation. We use the term 'cluster' to refer to the pre-existing allocations of, for example, classrooms. The simplest cluster randomised trial involves:

- consideration of the sample size required
- randomising the clusters involved into two groups
- applying the intervention to one of the groups
- measuring the outcome for the clusters involved
- comparing the results and interpreting the differences.

Considerations of sample size and analysis are more complex for cluster randomised trials. If in doubt, it is recommended that you consult a research organisation experienced in running large cluster randomised trials for advice. Each of the above steps is considered here for the simplest design.

Consideration of the sample size required

Please refer to Chapter 6 for calculation of the achieved sample size. Bear in mind that not everyone may want to take part, so you may want to take a larger starting sample than you will eventually need. For cluster trials in educational research, the consent of the schools involved is required. In some cases, pupil consent or that of their parents would also be needed. The timing and wording of the request for consent is crucial. It should:

- be requested before randomisation so that only those consenting clusters are randomised
- contain a rationale for the trial and, in particular, an explanation of why randomisation is being used
- contain an account of why it is just as important to carry out the outcome measurement if selected in the control group.

Trials should be 'ethically neutral' in that we do not know whether the intervention being trialled improves, makes no difference or is detrimental. In reality, there is often a perception in schools that the intervention is good and the role of the control group may be perceived as onerous. After all, why would one run a trial of a programme if one didn't really believe it was 'better' in some sense? For this reason, incentives can be offered to the control group such as the intervention itself (but delivered after the trial). Similarly, if a trial involves many stakeholders that require 'buy-in', visit to explain the principle behind an RCT may pay dividends.

You can gain an impression of the number that are likely to consent from previous exercises by yourself or from colleagues, or, if the worst comes to the worst, by asking a few informally. Also, bear in mind, people saying they will take part and actually doing so, are not necessarily the same thing, so you should allow for dropout here as well.

Randomising the clusters involved into two groups

Once a list of consenting clusters is obtained, randomisation can be carried out at the cluster level as in Chapter 3.

Applying the intervention to one of the groups

The intervention is then administered to the clusters selected to be part of the intervention group.

Measuring the outcome for the clusters involved

Please refer to Chapter 3.

Comparing the results and interpreting the differences

There are many quite complex methods for analysing data of this kind including, for example, multi-level modelling. However, for relatively small-scale applied research of the kind considered here, such as can be carried out by a teacher or school for their own purposes, a simple t-test comparison **of the cluster means** will often be sufficient.

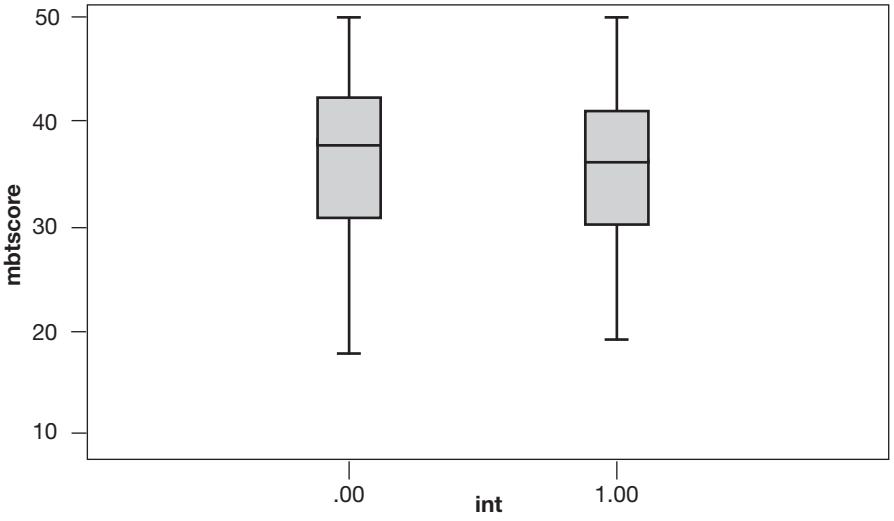
To analyse the results, we have to compare the average scores for the clusters (schools, classes, or whatever). As an example, we can consider a study in which schools were randomly allocated to a treatment or a control cluster. This is the same data that was used for Chapter 4. We now see that the analysis in Chapter 4 was problematic, since it ignored the clustered nature of the data. If individuals, rather than clusters, had been randomised, the Chapter 4 analysis would have been sound.

We start by showing a box and whisker plot of the clustered data (Figure 5.1). This can be obtained using the following SPSS syntax, where CONTACT is the school identifier:

```
aggregate outfile=*/break=contact/
matscore=mean(atscore)/mbtscore=mean(btscore)/int=
first(int).

examine vars=mbtscore by int/
plot=boxplot.
```

Figure 5.1 A box-and-whisker plot of the post-intervention measure for intervention and control groups (school means)



It looks slightly different from the plot in Chapter 4 because it is of school mean values. Please see Chapter 3 for a brief consideration of how to deal with missing data, and Chapter 7 for more detailed consideration. It can be seen from the plot that there does not seem to be much difference in outcome scores between the two groups: the median and inter-quartile ranges are very similar. The post-intervention measure using the school mean baseline scores as a background variable in a regression is analysed using the following syntax:

```
regression vars=matscore mbtscore int/
 dependent=mbtscore/
 method=enter.
```

Table 5.1 shows the result obtained (see Chapter 4 for an explanation).

Table 5.1 Coefficients

Model		Understandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std Error	Beta		
1	(Constant)	9.741	1.247		7.811	.000
	matscore	.803	.035	.921	22.654	.000
	int	.844	.651	.053	1.298	.197

We can conclude from this that the intervention does not have a significant effect (p=0.197). If we compare the result here with that for the unclustered analysis presented in Chapter 4, we see that the standard error of the coefficient for INT has increased from to 0.443 to 0.651. This is an illustration of why it is important to analyse clustered data in this way; by analysing the data at pupil level, we had underestimated the error associated with the intervention effect. In the example here it does not make a difference, but in other situations it could easily make the difference between a result being statistically significant and not.

'Splitting' clusters

One viable strategy is to sample schools and then randomise eligible pupils within each school. This is applicable to an intervention where small numbers of pupils are removed from classes to receive the intervention. This is very attractive since it improves statistical power as it is effectively an individually randomised trial and individual pupil scores can be analysed. However, care should be taken when randomising because schools vary in size. If a fixed number of eligible pupils are randomised to the intervention group in each school and the remainder get allocated to the control group, bias is introduced since pupils from larger schools are more likely to end up in the control group. If the intervention is only available to a fixed number of eligible pupils per school, the safest way to embark on randomisation is:

- from the eligible pupils in the school, randomly sample twice the number that can receive the intervention

- randomly assign half of these to the intervention group and half to the control.

This is an example of when it would be tempting to leave the sampling and randomisation up to the school. This is generally not a good idea since there is widespread evidence of 'subversion bias' (Torgerson and Torgerson, 2008) when randomisation is left to anyone other than the organisation running the trial. The term 'subversion bias' makes it sound like it is intentional, which sometimes it is; however, often there is just a lack of understanding about how to conduct randomisation.

6 Calculation of sample size

For the simplest design of a trial, where we are taking a random sample from a population of individuals with no baseline data, please refer to Chapter 3. Designs often require greater complexity than this. This chapter starts with some general considerations, including some basic sampling theory, and then addresses two common scenarios specifically: clustered samples and baseline data.

How big a sample do we need to design an experiment scientifically?

We assume that it has been decided to administer the intervention at school level. It would be simpler if we were able to take just some of the institutions involved, rather than the whole population: most reputable studies do this. Why is a sample satisfactory, rather than taking the whole population?

If we wanted to find the average height of a member of the UK population, we might select one individual at random, assuming that it was possible to do so, and measure their height. It is clear that this would not provide a very good estimate: if we had selected another individual, we might have got a completely different result. Now, if we were a race of invading Klingons, it might be enough to know whether the leading inhabitants were (say) the size of ants, humans or whales, but for practical use we would want a much more precise estimate. If, somehow, we could measure all inhabitants then there would be no sampling error. But, in addition to being expensive and difficult, it would almost certainly be overkill, in the sense that we are unlikely to need to know the result to such a degree of precision. Thus, a sample of around 7000 individuals gives a standard error for mean population height of approximately 0.1 cm, which is likely to be precise enough for most purposes. For this reason it is usual to take a sample, rather than the entire population.

To decide on the size of sample required, we can think about it in this way: 'How big a sample is it necessary to have in order to be **adequately sure that we can detect an effect** that is **large enough to be of educational**

significance and what allowance do we need to make for the fact that we are **dealing with clusters rather than individuals**?' This definition as it stands raises as many problems as it solves. What do we mean by 'educational significance', 'detect an effect', and 'adequately sure' and 'allowance for clustering'? We consider each of these in turn. The availability of good covariates (such as the baseline data discussed in Chapter 4) can have a substantial effect on the design of an experiment, specifically on the sample size required, and we consider this separately.

Educational significance

To identify significance, we have to think of a commonly acceptable way of measuring the size of an effect. So, for example, in other fields we use measures such as centimetres and currency to assess effects. This is not as easy as one might think at first sight in education or the social sciences.

In dealing with UK public exams, most readers would be prepared to accept an impact measured in terms of number of GCSEs, Highers or in exam grades. However, this type of widely agreed measurement is the exception rather than the rule. It could be more difficult to reach an intuitive agreement on what would constitute useful progress in terms of number of items correct on a reading test, or some kind of measure of behaviour. The widely used alternative is to compare the size of an effect with the population standard deviation. In a normally distributed population two-thirds of the population lie within one standard deviation (plus or minus) of the mean, and 95 per cent within two standard deviations. This is used to create a measure called effect size (ES), that is the fraction of the population standard deviation (SD) that a particular difference represents:

$$\text{ES} = \frac{\text{Size of difference}}{\text{Population SD}}$$

Even once we have reduced impact to terms of effect size, we still need to know how big is 'big' and how small is 'small'. Many social scientists follow Cohen (1988), whose conventional criteria **small, medium and large** are near ubiquitous across many fields. For Cohen an effect size of 0.2 to 0.3 might be a 'small' effect, around 0.5 a 'medium' effect and 0.8 to 1.0 a 'large' effect. However, it is important to recall that the terms

small, medium, and large are relative, not only to each other, but to the area of behavioural science, or even more particularly, to the specific content and research method being employed in any given investigation (Cohen, 1988). They should be used only as a last resort when one cannot think of anything else.[1] It is arguable that the definitions of small, medium and large do not carry over to education. For example, Wiliam (2008) suggests that 0.3 standard deviations is approximately the average progress in achievement made by a pupil in a year. In such a case, an effect size of 0.2 would be described as small using the Cohen conventions, but in fact would probably be considered rather large in an educational context. Wiliam suggests an effect size of 0.1 or even 0.05 as being substantively useful. The decision on what effect size to use is not cut and dried, and will require some thought depending on the circumstances and aim of the experiment.

Adequately sure that we can detect an effect

Two hypotheses are considered: the **Null Hypothesis (H_0)**, that there is no difference between the two groups, and the **Alternative Hypothesis (H_1)**, that there is a difference *d* between the two groups. The Null Hypothesis is unspecific (no difference) while the Alternative Hypothesis specifies an alternative value for *d* the difference.

The most widely used approach to statistical significance testing defines in advance the probability of making a **Type I Error**, that is, believing that we have detected an effect if, in fact, there is not one present. In statistical jargon, this is known as the **size** of the test, and it is usually set at 0.05. Using this value, if you repeated the experiment many times in the absence of a genuine effect, you would on average find one apparently statistically significant result in every 20 attempts.

Another aspect is what is called a **Type II Error**. This is, in a way, the inverse of the Type I error, that is, you believe that you have not shown an effect, when in fact there is one. In statistical jargon, this is related to the **power** of the test. The power of a test depends on the definition of the effect that you want to be able to detect. The probability of detecting

[1] Effect sizes with confidence intervals should also be reported, for use in meta-analyses.

an effect given that it exists is described as the **power** of a test. While the cut-off probability level for H_0 is conventionally 0.05, the conventional probability level for H_1 is 0.80, which is the power.

One other question to be considered is whether a one-sided or two-sided test is to be used. This is a tricky question. Historically and conventionally, two-sided tests are used. That is to say, we test not just whether the intervention group performs better than control, but also whether it performs worse. The alternative is called a one-sided test, which tests differences in just one direction, say just that the intervention group is better. In many ways in this context a one-sided approach makes more sense. Suppose you were considering introducing a new treatment. Your decision space would be:

a) new treatment certainly better: introduce it

b) new treatment not certainly shown to be better: forget it.

In many cases all you really want to know is whether a new treatment is better than the old one: you are not really interested in whether it is worse. If it is not an improvement, then forget it. So a one-sided test could make more sense. A one-sided test would have the benefit that it was more likely to detect effects. Unfortunately there is such a strong convention that two-sided tests are used, that it would probably not be acceptable to use a one-sided test. If you want to publish your results in a journal, editors might be reluctant to accept an article with one-sided significance and there might a suggestion of 'what are you hiding?' Whatever is decided, it is important to specify in advance if results are going to be analysed using a one-sided test.

Allowing for clustering

One other aspect of a design is clustering effects. These weaken both size and power. The standard method of assessing an effect, namely the use of a **t-test** on individual test scores, does not apply in this situation because the data is clustered, and, as already noted, it is as if we have a rather smaller sample.

What allowance do we need to make for clustering? This depends on how alike the subjects are within the clusters. If everyone were the same

within a cluster at the end of the study, then there would be little point in taking more than one individual in a cluster. If, on the other hand, there were no relation between which cluster an individual is in and other characteristics, this would be the same as allocating at individual level.

Before going any further about determining sample size, we define another two statistics: the intra-cluster correlation ρ, and the design effect (Deff). The degree of similarity is indexed by the statistic ρ, which gives the ratio between the cluster variance and the total variance:

$$\rho = \frac{\text{Between variation}}{\text{Total variation}}$$

The design effect (*Deff*) is the extent to which the size of the sample needs to be increased to achieve the same precision as a simple random sample and is defined as:

$$\text{Deff} = 1 + (b - 1)\,\rho$$

where b is the mean cluster size.

Statisticians working in designing randomised trials frequently think in terms of the Minimum Detectable Effect Size (MDES). The MDES varies according to the **size** (P-value for rejection) and **power**, but a convention has been widely adopted that the probability of a Type I Error (rejecting the Null Hypothesis when it is true) is taken as the standard 0.05 and the power as 0.80. The MDES is the smallest effect that the researcher would have an 80 per cent chance of detecting under these conditions.

Before undertaking a study, the researcher should specify how small an effect they would want to detect. The design should be able to detect an effect of at least this size. An approximate formula for MDES is given by Bloom *et al.*, 2007:

$$\text{MDES} = M\sqrt{\frac{\rho}{P\,(1 - P)\,J} + \frac{1-\rho}{P\,(1 - P)\,n\,J}}$$

ρ is the intra-cluster correlation (without covariates), that is the proportion of the overall variance due to variation between the clusters.

P is the proportion of clusters (for example, schools) allocated to the treatment group by the randomisation procedure.

n is the number of individuals (for example, pupils) in each cluster.

J is the total number of clusters randomised.

M is a multiplier based on the t-distribution.

There are two terms inside the square root sign. Both are divided by the factor $P(1-P)$. This reaches a maximum when P=0.50, showing that the most effective design is when the allocation between treatment and control is on an equal basis. The first term under the radical,

$$\frac{\rho}{P\,(1 - P)\,J}$$

relates to the variation between clusters, with ρ proportional to the between-clusters overall variance, and the divisor J, relating to the number of clusters. The second term under the radical,

$$\frac{1-\rho}{P\,(1 - P)\,n\,J}$$

relates to the variation within clusters, with $1-\rho$ proportional to the within-clusters overall variance, and the divisor nJ, relating to the number of individuals. Since the divisor for the second term is the product of that for the first term and n, it is going to be substantially larger and so it will generally happen that the between-cluster variation is more important than the within-cluster variation in determining MDES.

Example 1

Suppose we want to be able to identify with confidence an effect size of 0.5 and the intra-cluster correlation ρ is 0.2. Will a sample of 40 schools, 10 pupils in each, be sufficient? We assume that the allocation ratio is 0.5 to treatment, 0.5 to control. The value of M is approximately 2.8:

MDES = 2.8*sqrt[0.2/(0.5*0.5*40)+(1-0.2)/(0.5*0.5*10*40)]

=2.8*sqrt[0.02+0.008]

= 2.8*0.167

= 0.47

Since the calculated size is smaller than the stated effect size of 0.5, under this design we are able to detect the specified size, so the design is satisfactory.

Deciding on suitable values for the MDES formula

MDES

This was discussed in the section on educational significance, and to some extent it is a matter of judgement for those involved. The possible benefits of the innovation have to be balanced against the costs, either in time and effort, or in actual financial terms, and also in terms of what is realistic.

To detect a really small effect may require a sample larger than the researcher considers feasible. In such a case the researcher has to consider whether the number involved in the experiment should be increased, perhaps bringing in more participants, or whether the focus should be only on detecting relatively large effects, or even whether the plan as it stands should be reconsidered.

This is important: there will be no point in deciding that you would be interested in an effect of one-tenth of a standard deviation, and designing an experiment that could not detect it.

Multiplier, *M*

This is a number that depends on the size of the test, the power desired and whether a one-tailed or two-tailed test is used. As discussed above, there is a convention of size = 0.05 and power = 0.80. Most experimenters would want to follow this, especially if they want to publish the results. When J exceeds about 20, the value for this multiplier is approximately 2.8 for a two-tailed test and 2.5 for a one-tailed test, given size = 0.05 and power = 0.80 (Bloom, 1995).

Allocation proportion, *P*

The most efficient allocation statistically would be 50:50 between experimental and control group. However, it may be that it is easier or cheaper to collect the data for one group. For example, in a school setting it could be that there is no special input to the control group and the data required is collected anyway in the course of normal school work. In this situation it would make most sense to increase the number in the control group.

Intra-cluster correlation, ρ

This is where things start to get a little tricky. The experimenter appears to be caught in a catch-22. To carry out the study you need to know this information; to find out the information you need to carry out the study. How does one get this information without actually conducting the study? If this is your first study in this type of area, this might indeed be difficult. However, there are ways around this.

- A study of the background literature might reveal results that give an indication. If the information is not available in an article, and a publication is relatively recent, it may be worth contacting the authors directly.

- Some articles have been published specifically looking at intra-cluster correlations. For example, Hedges and Hedberg (2007) have produced extensive tables of intra-cluster correlations by age for reading and mathematics. These are for the USA, but will at least provide a ballpark estimate. Hutchison (2009) provides a selection of values of intra-cluster correlations within schools by a range of wide topic areas, such as leisure activities and attitudes to school.

- If all else fails, then, for schools, most academic attainment variables have a value of ρ in the neighbourhood of 0.2, so assuming 0.25 will be a safe bet, and most attitudinal and lifestyle variables will have a value of ρ in the neighbourhood of 0.05 or less. Be aware that if you sample entire classes in streamed schools, then intra-cluster correlation for attainment variables is likely to be very high.

Number in each cluster, *n*

While the number of respondents within each cluster has an effect on MDES, it is less important than the number of clusters: 20–30 is often a useful number of subjects within each cluster. Because the precise number is less important, it can often be useful to allow practical considerations to have a substantial impact. For example, it will be relevant to consider the number of interviews one can do in a day or a morning. Or, if pupils are withdrawn from normal classes to form an experimental group, then the number that can be taught in such a format will be important.

Number of clusters randomised, *J*

Finally, the number of clusters (for example, schools) is a very important determinant of the MDES. This is what should be increased (or decreased) to ensure an appropriate MDES is obtained.

Example 2

A senior civil servant wishes to investigate the effect of a new method of teaching reading. She would like to use an experimental design to ensure that findings cannot be undermined by the use of self-selected samples. Even a quite small improvement would be educationally important over the whole country, so she decides to select an MDES of 0.2. Background literature suggests that the intra-cluster correlation will be around 0.2. Will a sample of 100 schools, with one class of 25 pupils selected randomly in each school from the target year group, be sufficient for this purpose? We assume that the allocation ratio is 0.5 to treatment, 0.5 to control.

MDES = 2.8*sqrt[0.2/(0.5*0.5*100)+(1-0.2)/(0.5*0.5*100*25)]

= 2.8*sqrt[0.008+0.00128]

= 2.8*0.0963

= 0.27

Since the MDES from this design is larger than that sought, it is not going to be satisfactory. As an exercise, the reader might like to show that around 180 schools would be required.

Baseline data

Another aspect is the use of covariates such as the baseline data discussed in Chapter 4. Such data could also consist of pupil test scores on a different test or the attitude of the people enrolled in a campaign at the start of the study, or some kind of aggregated measure, such as the proportion of pupils with 5+ A*–C grades at GCSE.

In the previous section, we discussed how to determine the MDES for a completely randomised cluster design. It was emphasised that before undertaking a study, the researcher should specify how small an effect they would want to detect, and if the suggested design gives an MDES of greater than that, the design is not strong enough.

Cluster randomised trials tend to need substantially more participating individuals because individuals within the same cluster tend to be more similar than individuals in different clusters; and the power of the design is very much driven by the number of clusters, much more than by the number of individuals within clusters. It can be that to get a sufficiently powerful design to detect the type of effect that can be of interest in education, we need to increase the number of clusters to a prohibitive number. Fortunately, there is another way in which we can increase the power of a cluster-level RCT, and this is by using covariates.

Do we need these covariates to be at the cluster or individual participant level? For this discussion it is helpful to term the individual participant as 'level 1' and the cluster as 'level 2'. Intuitively one might expect that covariates would be 'stronger' at level 1 since there is more variance at this level. However, the situation is different in this case, since we are investigating effects at level 2. In actual fact, level-1 covariates can be expected to have an effect on level-2 differences, since their aggregated means will differ between clusters. Also they can be expected overall to have a stronger effect than level-2 covariates since there should also be an effect on within-cluster variation. Although there are more sophisticated ways of analysing clustered data, we are proposing in Chapter 5 that cluster means are analysed. For this reason, level-2 covariates would be sufficient here.

In order to calculate the sample size needed for a cluster randomised trial with both level-1 and level-2 covariates, an approximate formula can be used (Bloom *et al.*, 2007):

$$\text{MDES} = M_{J\text{-}K}\sqrt{\frac{\rho(1 - R^2_C)}{P\,(1 - P)\,J} + \frac{(1 - \rho)\,(1 - R^2_I)}{P\,(1 - P)\,n\,J}}$$

ρ is the intra-cluster correlation (without covariates), that is the proportion of the overall variance due to variation between the clusters.

P is the proportion of clusters (for example, schools) allocated to the treatment group by the randomisation procedure.

n is the number of individuals (for example, pupils) in each cluster.

J is the total number of clusters randomised.

K is the number of level-2 covariates included in the model.

M_{J-K} is a multiplier based on the t-distribution.

R^2_C is the proportion of the random variation between schools that is reduced by the covariates.

R^2_I is the proportion of the random variation within schools that is reduced by the covariates.

If there are no covariates then both R^2_C and R^2_I are equal to zero and the formula reduces to that in the previous section (see equation on p.43).

Further consideration of this formula is similar to the previous section. There are two terms inside the square root sign. Both are divided by the factor $P(1-P)$. This reaches a maximum where P=0.50, showing that the most effective design is when the allocation between treatment and control is on an equal basis. The first term under the radical,

$$\frac{\rho\,(1 - R^2_C)}{P\,(1 - P)\,J}$$

relates to the variation between clusters, with ρ proportional to the between-cluster variance and $(1 - R^2_I)$ the proportion of this remaining after the effect of the covariates. The second term under the radical,

$$\frac{(1 - \rho)\,(1 - R^2_I)}{P\,(1 - P)\,n\,J}$$

relates to the variation within clusters, with $(1-\rho)$ proportional to the within clusters overall variance, $(1 - R^2_I)$ the proportion of this remaining after the effect of the covariates. Since the divisor for the second term is the product of that for the first term and n, it is going to be substantially larger and so it will generally happen that the between-cluster variation is more important than the within-cluster variation in this design stage. This implies in turn that the important aspect of any covariate is the amount of between-cluster variation it explains.

Example 3

Suppose, as before, we want to be able to identify with confidence an effect size of 0.5 and the intra-cluster correlation ρ is 0.2. We also know that the proportions R^2_C and R^2_I are both 0.5. Will a sample of 40 schools, 10 pupils in each, be sufficient? We assume that the allocation ratio is 0.5 to treatment, 0.5 to control. The value of $M_{J\text{-}K}$ is approximately 2.8 (see below).

MDES=2.8*sqrt[(0.2*(1-0.5)/(0.5*0.5*40)+(1-0.2)*(1-0.5)/(0.5*0.5*40*10)]

= 2.8*sqrt[0.01+0.004]

= 2.8*0.118

= 0.33

Since the calculated size is smaller than the stated effect size of 0.5, under this design we are able to detect the specified size, so the design is satisfactory. We would expect this, since the design without covariates was already satisfactory, and this is a more powerful design. Under this design, however, we could substantially reduce the sample size and, thus, save money, or we could detect a smaller effect.

Deciding on suitable values for the MDES formula

Many of the variables in this formula are the same as those listed in the previous section. New variables are listed here.

Multiplier, $M_{J\text{-}K}$

As in the previous formula, this depends on the 'degrees of freedom'. In the case of no covariates, its value can be decided on the basis of the number of clusters randomised, J. Here, we need to look at (J-K), the difference between the number of clusters and the number of level-2 covariates. If (J-K) exceeds about 20, the value of the multiplier is approximately 2.8 for a two-tailed test and 2.5 for a one-tailed test, given power=0.80 and size=0.05 (Bloom, 1995).

R^2_C and R^2_I

If values of ρ can be difficult to find, R^2_C and R^2_I are likely to be even more so. We can get an estimate of these in background literature and Bloom *et al.* (2007) give some results for elementary, middle and high schools in the USA.

Since the major contribution to this formula comes from the cluster covariate effect, one can get a reasonable estimate simply from trying to determine this. Even if one assumes that R^2_I is zero, the estimate is not much weakened. This approach is applicable to the analysis of clustered data proposed in this guide: using cluster means. Thus, in the previous example, the MDES is now given by:

MDES=2.8*sqrt[(0.2*(1-0.5)/(0.5*0.5*40)+(1-0.2)*(1-0)/(0.5*0.5*40*10)]

= 2.8*sqrt[0.01+0.008]

= 2.8*0.134

= 0.38

Looking at schools, Bloom *et al.* (2007) show that earlier years' aggregated data can be a readily available and relatively valid source of estimates of R^2_C . A ballpark estimate could be obtained from the square of the correlation of equivalent baseline and post-test scores for the age group of interest.

It is emphasised that estimates of R^2_C, R^2_I and ρ are likely to be somewhat ad hoc. It will be desirable to leave a margin on any MDES thus obtained. However, even the use of ad hoc estimates is likely to be better than ignoring these crucial aspects of a cluster randomised trial.

7 Dropout: prevention is better than cure

It is helpful here to categorise types of dropout (Shadish *et al.*, 2002).

- **Measurement attrition** refers to a failure to complete outcome measurement, whether or not treatment is completed.
- **Treatment attrition** refers to those research participants who do not continue treatment, whether or not they continue with the measurement protocol.

The best way of coping with dropout is to avoid it: prevention is better than cure. Furthermore, it is better to prevent measurement attrition even when you cannot prevent treatment attrition. This is because an intention-to-treat analysis should be used regardless of the extent of treatment attrition (Torgerson and Torgerson, 2008). Addressing the issue of measurement attrition and, specifically, whether dropout is treatment correlated, is more problematic and beyond the scope of this guide. Shadish *et al.* (2002) provide an introduction to the issues with appropriate further references.

Strategies to prevent dropout

To take part in a properly executed experiment can represent quite a serious commitment on the part of those involved. There are likely to be self-study or training sessions, time commitments, change of one's ordinary practice, testing (perhaps before and after) record-keeping and an unwonted (and probably unwanted) degree of supervision of one's work. All of these are likely to deter interest in actually carrying out a project, and, even where signed up in the first instance, to continuing to perform the tasks involved. One can imagine that this is even more likely to occur for those in the control group, who do not even have the excitement of implementing the new procedure. So how do we stop such subjects dropping out?

It is all too easy to think in terms of the experimenter (or the experiment team) on the one hand, and the subjects on the other. Taking part in an experiment represents a substantially greater commitment than filling in a survey, generally, and it is important to take account of this. Participants may have an interest and a degree of expertise in the topic, and a commitment to doing things as well as possible. Thus it would be valuable to have input from them in designing the intervention. Small infelicities in the instructions can undermine important parts of an intervention, and those familiar with the field can spot these before they ruin the project.

The typical response of a profession, presented with a possible new 'magic potion', will be either to demand that it is given to everyone or to those most in need. The research team has to convince those likely to be involved that it is worth taking part, and continuing. Crucially it also means that the experimenter and participants all agree on the value of the experimental approach, and that they will all agree to cooperate even if allocated to the unglamorous control group. This requires an understanding of the principles of randomisation (at least by the teachers involved, say) in order that they understand the reason behind this random denial of the intervention.

It is, therefore, often appropriate to visit centres involved in the trial in order to explain the methodology and obtain their support for the study. Part of this explanation could include an account of the **equipoise** principle. This holds that a subject may be enrolled in an RCT only if there is true uncertainty about which of the trial arms is most likely to benefit them. In some ways this is rather a counter-intuitive requirement. What would be the point of trying a new technique if you did not believe it was going to be helpful? The target audience is not just the person who devised the scheme. Their colleagues must be convinced and, potentially, education policy-makers. Even when experts agree that an innovation is going to be beneficial, it can happen that it simply is not. Muir (2008) quotes the cases of school driving lessons, Scared Straight programmes and other initiatives which despite being 'obviously beneficial' in advance, actually appeared to worsen the situation when trialled.

Those taking part in an experiment should, if at all possible, take part because they believe in it, rather than for any other reasons. That said,

there are stratagems and inducements which can help increase and maintain cooperation.

- Ensure that randomisation happens after consent has been given. For a trial where schools are randomised, agreement to participate can be as low as 11 per cent (Smith *et al.*, 2007). In the same study, 94 per cent of schools that agreed were still on board by the end of the study and returned completed post-intervention tests. Dropout was hence a minor problem since randomisation of consenting schools was carried out.
- A delayed introduction approach can be used where one group gets the initiative at the start and after the initiative is complete, the control group also receives it. In practice, it is often necessary to 'reward' the control group in this way to ensure that headteachers signing up to the trial perceive some benefit to involvement. It could be that the initiative shows no benefit or is detrimental, in which case such a plan is problematic.
- Schools could be offered help in coping with the assessments. Thus the project could offer to pay for supply teacher cover for teachers involved in administering assessments.
- Keep schools involved. A full and reasonably timely description of pupil scores is likely to be of interest to anyone taking part and this can be supplied to schools at the end of the trial.

Lastly, it should be noted that the implementation of the intervention should mimic a real-life scenario. It is no good having some enthusiastic proponent of the intervention materials phoning up schools to persuade to use them properly; we are interested in how the intervention will function in the real world.

References

Bloom, H.S. (1995). 'Minimum detectable effects: A simple way to report the statistical power of experimental designs', *Evaluation Review*, **19**, 5, 547–556.

Bloom, H.S., Richburg-Hayes, L. and Rebeck Black, A. (2007). 'Using covariates to improve precision for studies that randomise schools to evaluate educational intervention', *Education Evaluation and Policy Analysis*, **29**, 1, 30–59..

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

Cook, T.D. (2002). 'Randomized experiments in educational policy research: a critical examination of the reasons the education evaluation community has offered for not doing them', *Educational Evaluation and Policy Analysis*, **24**, 3, 175–199.

Cook, T.D. and Payne, M.R. (2002). 'Objecting to objections to random assignment.' In: Mosteller, F. and Boruch, R. (Eds), *Evidence Matters – Randomized Trials in Educational Research*, First edn. Washington D.C.: Brookings Institution Press..

Heckman, J.J., Ichimura, H. and Todd, P. (1997). 'Matching as an econometric evaluation estimator: evidence from evaluating a job training programme', *Review of Economic Studies*, **64**, 4, 605–654.

Heckman, J.J. and Smith, J.A. (1995). 'Assessing the case for social experiments', *Journal of Economic Perspectives*, **9**, 2, 85–110.

Hedges, L.V. and Hedberg, E.C. (2007). 'Intraclass correlation values for planning group-randomized trials in education', *Education Evaluation and Policy Analysis*, **29**, 1, 60–87.

Hutchison, D.H. (2009). 'Designing your sample efficiently: clustering effects in education surveys', *Educational Research*, **51**, 1, 109–126.

Lehr, R. (1992). 'Sixteen S-squared over D-squared: a relation for crude sample size estimates', *Statistics in Medicine,* **11**, 1099–1102.

Muir, H. (2008). 'Science rules OK!', *New Scientist*, 2657, 40–43.

Oakley, A. (2006). 'Resistances to 'new' technologies of evaluation: education research in the UK as a case study', *Evidence & Policy*, **2**, 1, 63–87.

Oates, T. (2007). 'Protecting the innocent – the need for ethical frameworks within mass educational innovation.' In: Saunders, L. (Ed.), *Educational Research and Policy-Making – Exploring the Border Country Between Research and Policy.* First edn. Abingdon: Routledge.

Shadish, W.R., Cook, T.D. and Campbell, D.T. (2002). *Experimental and Quasi-Experimental Designs.* Boston: Houghton Mifflin Company.

Silverman, W. (1997). 'Risks and benefits of medical innovations.' In: Maynard, A. and Chalmers, I. (Eds) *Non-Random Reflections on Health Services Research.* London: BMJ Publishing Group.

Smith, P., Styles, B. and Morris, M. (2007). *Evaluation of Rapid Reading*, NFER: Slough.

Styles, B.J. (2009). 'The future is random – why the RCT should often be the method of evaluation.' In: St.Clair, R. (Ed.) *Education Science: Critical Perspectives.* First edn. Rotterdam: Sense Publishers.

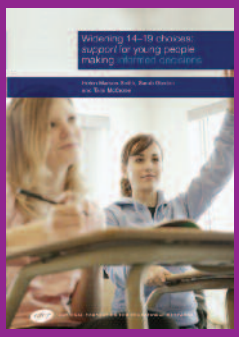
Torgerson, D. and Torgerson, C. (2008). *Designing Randomised Trials in Health, Education and the Social Sciences.* Basingstoke: Palgrave Macmillan.

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

Also available from NFER

1

2



3

4

1 The impact of 14–16 year olds on further education colleges
The central aim of this research was to examine the strategies that FE colleges and their staff used to integrate 14–16 year olds successfully into their institutions and to explore the impact that 14–16 year olds have on FE colleges, their staff and older learners.
www.nfer.ac.uk/publications/ICL01/

2 Widening 14–19 choices: support for young people making informed decisions
This is a summary of key findings from NFER's recent work relating to 14–19 education in order to understand better how young people of this age group are navigating their way through complex choices of qualifications and locations of study.
www.nfer.ac.uk/publications/SMD01/

3 Attitudes to reading at ages nine and eleven: full report
In June 2007 NFER ran a reading survey questionnaire to determine current attitudes to reading. The questions dealt with enjoyment of reading and confidence in reading. This report looks at the results of the questionnaire.
www.nfer.ac.uk/publications/RAQ01/

4 The value of social care professionals working in extended schools
As the collaboration between social care and education professionals develops, the question of what role social care professionals can take in school, as well as what should be expected of them, is becoming increasingly pertinent. This report looks at levels of integration and the role of social care professionals.
www.nfer.ac.uk/publications/SCX01/

A guide to running randomised controlled trials for educational researchers

Randomised controlled trials (RCTs) are seen as the gold standard for evidence-based educational practice. This guide examines when they should be used to evaluate an intervention and when other approaches may be more suitable.
It covers:

- differences and similarities between medical and educational experiments
- the simplest design for an RCT
- testing before an intervention
- clustered data
- calculation of sample size
- the issue of dropout and how to prevent it.

It is essential reading for educational researchers and those commissioning research.

National Foundation
for Educational Research
The Mere Upton Park
Slough Berkshire SL1 2DQ

T: 01753 574123
F: 01753 691632
E: enquiries@nfer.ac.uk
W: www.nfer.ac.uk

ISBN 978 1 906792 68 8

£8.00