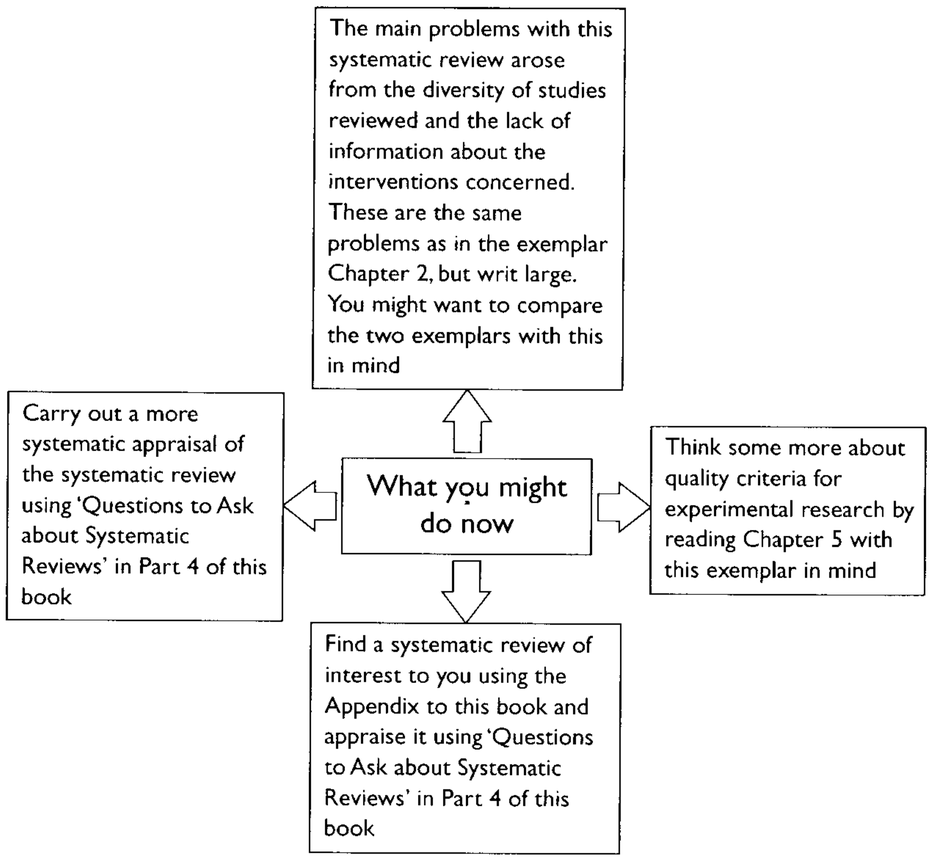
41 Nicol A.R., Stretch D.D., Davison I., Fundudis T. Controlled comparison of three interventions for mother and toddler problems: preliminary communication. *J R Soc Med* 1984; 77: 488–91.
42 Olds D.L., Kitzman H.J., Cole R.E. Effect of home visitation by nurses on caregiving and maternal life course. *Arch Pediatr Adolesc Med* 1995; 149: 76.
43 The Infant Health and Development Program. Enhancing the outcomes of low-birth-weight, premature infants. *JAMA* 1990; 263: 3035–42.
44 Dickersin K., Berlin J.A. Meta-analysis: state of the science. *Epidemiol Rev* 1992; 143: 154–76.

What you might do now



CHAPTER 5

THE BASICS OF EXPERIMENTAL DESIGN

Introduction

Experiments are particularly important in health care research. It has been argued that they should be more important in social care research too (Oakley and Fullerton, 1996). Some people claim that experimental methods are the only methods capable of investigating causality. They are certainly superior to all other methods in this regard. It is not possible to decide whether some health or social care intervention is effective if it is not clear what causes what effects. Thus the most telling evidence about effective practice is evidence that comes from experimental work.

The major problem in investigating causality is that everything that happens has multiple causes. A controlled experiment is an artificial situation established so that the multiple causes of phenomena can be controlled, by excluding some influences, standardising others, while allowing others to vary. This is described as *controlling variables to prevent confounding*, where 'confounding' means muddling the picture so that it is difficult to discern what is causing what to happen. The principle is much the same as that used by an electrician in isolating a circuit in a complex electrical system and then running various charges between different points at known amplitudes and seeing what happens. This chapter describes the way in which experiments are designed. Chapter 6 looks at the instruments which are used for collecting data in experimental research and Chapter 7 at the more common ways in which the results of experiments are expressed.

In health and social care research the terms 'experiment' and 'trial' are often used interchangeably, though the use of the term 'trial' usually implies that what is being investigated is the effectiveness of a health or social care intervention.

1 Experiments as systems of safeguards

Experiments usually involve treating two or more groups differently. This is often expressed by saying that an experiment has 'arms', each arm consisting of a group of subjects who have been subjected to different treatments.

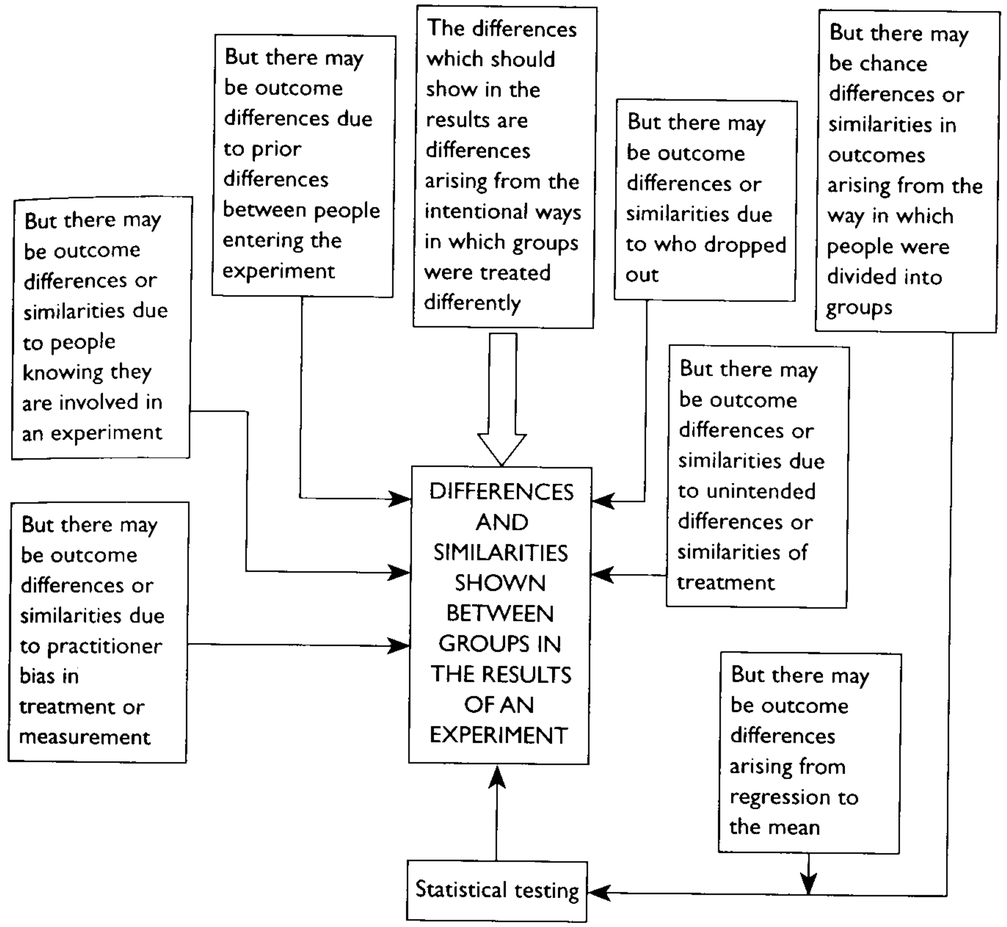
At the end of an experiment there will be some results which will show either that there is a difference in outcomes between the arms, or that there is not. Chapter 7 gives an account of some of the more usual ways in which the results of experiments are expressed. Any such difference may be statistically significant or not (see Chapter 7, sections 1 and 2). A *difference* in outcome between the arms of the experiment should show the effect of what was done differently and intentionally in different arms and nothing else except the effects of chance. Or *similarities* in outcomes should show that what was done differently and intentionally had much the same effect, and that except for the play of chance, there was nothing else creating this similarity. But unfortunately there are plenty of opportunities for the results of experiments to reflect matters other than the intended differences or similarities in treatment. Figure 5.1 shows the ways in which other factors may contaminate the results. Figure 5.2 tells the same story but identifies a number of safeguards which can be used to block off these routes of confusion.

2 Double-blinded, randomised controlled experiments or trials and other experimental designs

Figure 5.3 gives a picture of an RCT. The terms in it will be explained as the chapter proceeds. Figure 5.4 gives as an example the structure of the RCT that formed the basis for the economic analysis which is presented as the exemplar study in Chapter 3.

The randomised controlled experiment or trial (RCT), with double-blinding, is the experimental design which is regarded as the 'gold standard' in health research, particularly for testing the efficacy of drugs and other treatments. In terms of Figure 5.2, it contains the most stringent array of safeguards against confounding and against bias. Where it can be used, and used appropriately, it is certainly the design which produces the results that can be regarded with most confidence. Although it is often impracticable or unethical to use an RCT design, all other experimental designs can be regarded as

Figure 5.1 Routes of confusion in an experiment



deficient versions of this format, lacking one or more of the characteristics which give the RCT its power to investigate causality. Other experimental designs are often called 'quasi-experiments', the 'true' experiment being the kind involving randomisation.

3 Forming comparison groups – creating controls

(See boxes 1 and 6 on Figure 5.2)

An RCT (Figures 5.3 and 5.4) starts with a sample of people drawn from a wider population. How the sample is drawn, and whether it is representative of a wider population, and what wider population it is representative of, is important for the question of how far the results can be generalised (see section 12 later). Matters other than representativeness will also determine who enters the trial, such as estimations of whether people would be harmed by participating, and individual choice as to participate or not (see Figure 5.4). Once a pool of subjects has been assembled they are divided *at random* into as

Figure 5.2 Safeguards against misleading results in a controlled experiment

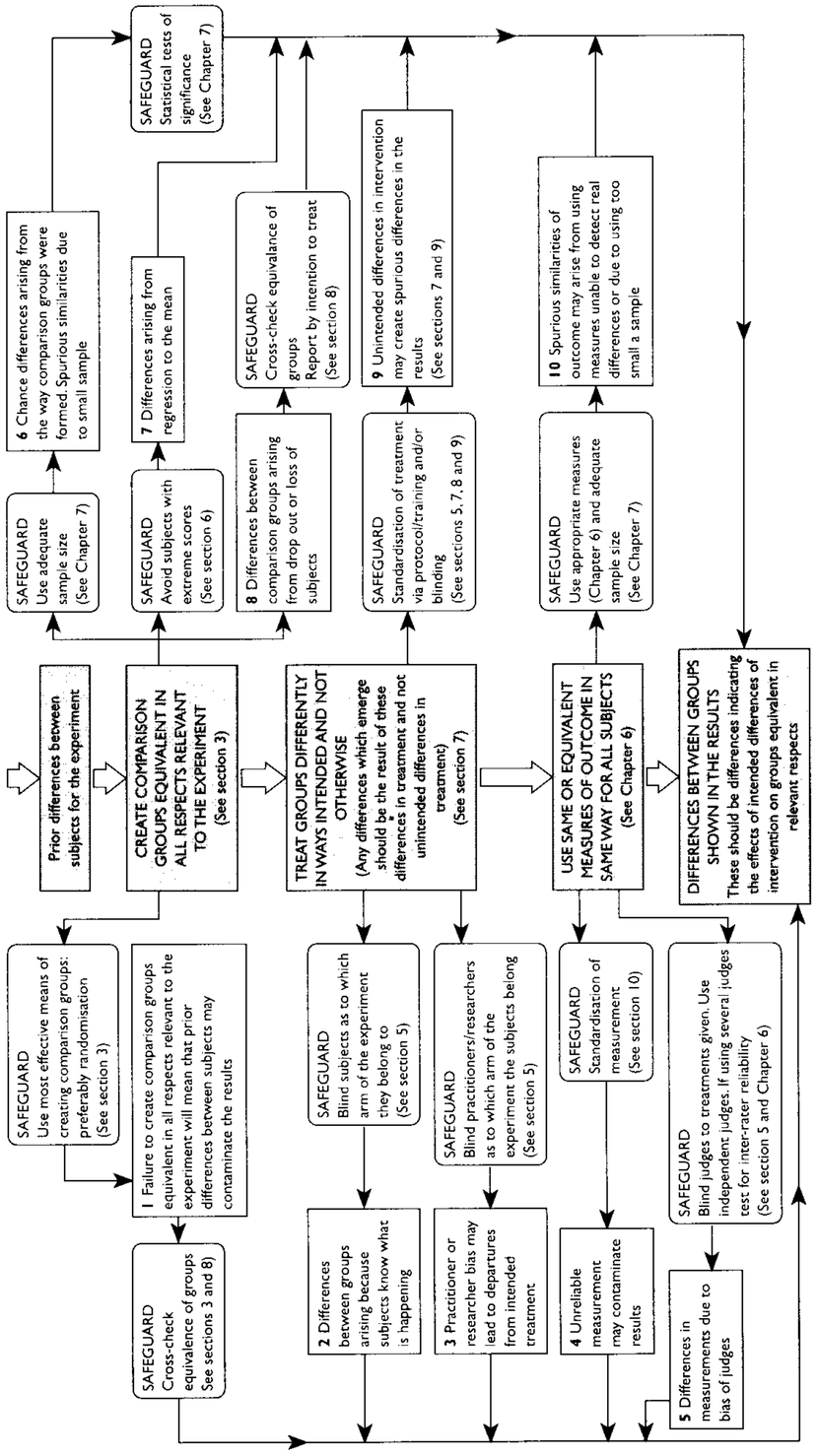


Figure 5.3 A basic randomised controlled trial

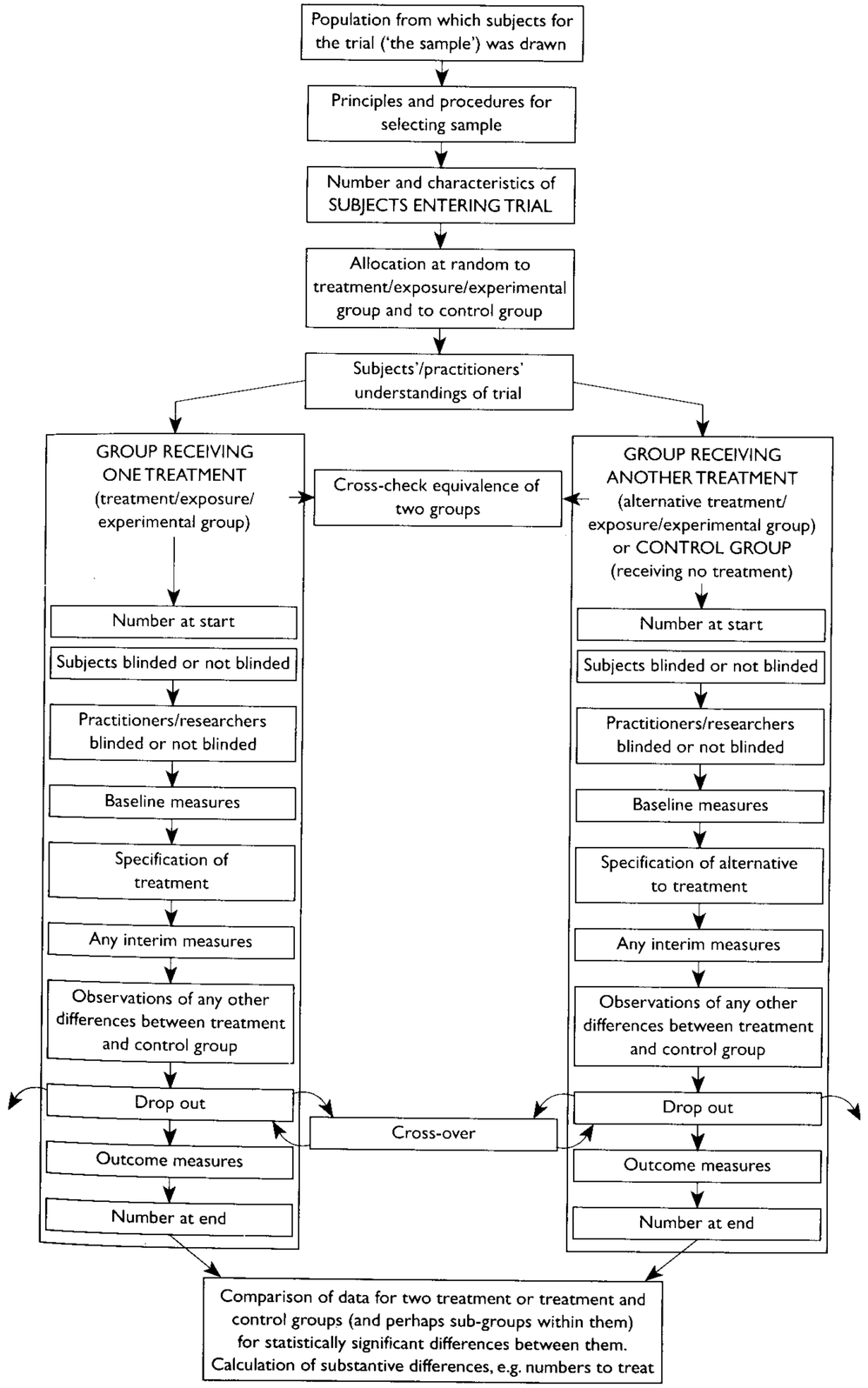
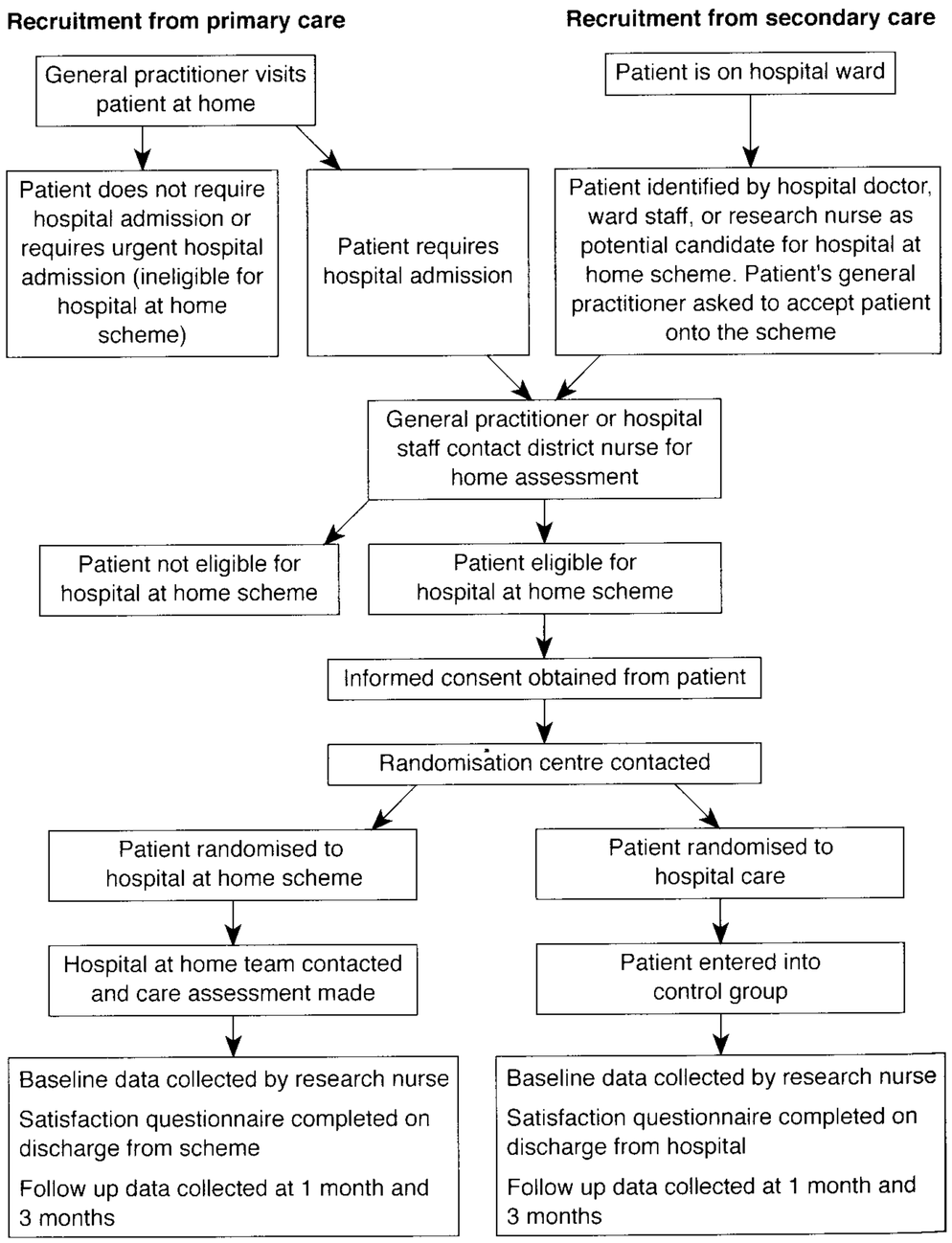


Figure 5.4 Procedure for recruitment and randomisation of patients and data collection in a randomised controlled trial comparing the relative effectiveness of a hospital at home scheme with inpatient care (Shepperd et al., 1998: 1787)



many groups as the trial has arms. The trials in Figures 5.3 and 5.4 have two arms. Random allocation might be done simply by tossing a coin for each subject. Thus an RCT involves two (or more) random samples, which together add up to 100 per cent of the subjects involved.

Using randomisation to create comparison groups is the most effective way of creating groups that are similar. While groups can be created in other ways (see Box 5.1), most of the alternatives rely on dividing people according to known differences. So long as the sample is large enough, randomisation should also produce comparison groups that are similar with regard to unknown differences as well. Schultz and his colleagues (1995) suggest that failure to randomise properly can spuriously inflate outcome differences between arms of a trial by up to 40 per cent. Using methods for creating comparison groups other than randomisation may result in the entirety of any outcome differences between arms in an experiment being due to pre-existing differences between subjects, rather than due to the different ways subjects were treated.

Randomisation may need a helping hand to distribute the characteristics of subjects evenly across arms of the trial. For example, if age were an important variable a researcher might first divide the pool of subjects into age groups and then randomly allocate the members of each age group to arms of the trial, thus ensuring that each arm had a similar age profile (Bowling, 1997: 214–17).

Randomisation also excludes the possibility of subjects, practitioners or researchers influencing which treatment subjects are allocated to and hence biasing the results.

Cross-checks can be made to see whether randomisation has created similar groups in terms of known characteristics. For example, there should be nearly equal numbers of males and females in each group, or each group should have a similar age profile.

Box 5.1 gives a range of methods used for creating comparison groups.

4 Sampling units

In most experiments the sampling unit is an individual person and in an RCT it is usually individuals who are randomly allocated to different interventions. However, sometimes the units might be parts of people, clinics, or communities. The experiment reported in Chapter 1 compares the effects of two different bandaging systems on the healing of venous leg ulcers. In that experiment it was not individuals that were allocated to the two different bandaging systems, but ulcerated legs. Thus the same person might appear twice in the results, once for each leg. The sampling units for an experiment should be decided, and the experiment arranged, such that what happens to any one unit is independent of what happens to any other. Thus in an experiment concerning group therapy, what happens to one person in a therapy group will be influenced by and in turn influence what happens to others. There is a case for saying that a therapy group of ten

Box 5.1 Some ways of attempting to create comparison groups which are similar in all relevant respects*

Randomised controlled trials

Subjects already selected to be rather similar to each other are allocated at random to separate groups: to a control group and one or more experimental/treatment group(s) *or* to two or more groups receiving different interventions (see Figures 5.3 and 5.4). Randomisation may be *unrestricted* or *stratified*. The latter means that the initial pool of subjects is first divided into categories (age groups, genders, ethnic groups) and each subject in each of these groups is then is randomly allocated (Bowling, 1997: 214–17).

Cross-over designs

The same people feature as both experimental and control subjects. For example, the subjects for the experiment are divided into two groups. One group gets the placebo first (see section 5) and then the drug being tested and one group first gets the drug being tested, and then the placebo. Where this design is used it is often as an additional feature of an RCT, with random allocation to groups. And sometimes an RCT will use four groups (or more): *c* then *c*, *c* then *e*, *e* then *c* and *e* then *e* (*c* = control, *e* = experimental group). Where this is possible this is an extremely powerful design, but the opportunities for using it are limited to situations where the intervention received has no long-lasting effect (Roberts and Sibbald, 1998). N-of-1 experiments and single case evaluations are special versions of cross-over designs (see section 13).

Matched pairs designs

Subjects for the experiment are paired off to create pairs of people who are similar to each other in ways considered relevant to the experiment. One of each pair is allocated to one arm and the other to the other arm of the trial – usually at random (Bowling, 1997: 220–2); though this is usually described as 'quasi-randomisation'. Sometimes matched pairs designs are attempted retrospectively where 'control pairs' are found for subjects of an experiment which has already been done. This is one approach used in 'case control studies' (see Chapter 10).

Purposive sampling, sometimes called 'factorialisation'

Two groups are created which are as near identical in all respects except one – for example, a group of males and a group of females with similar age, social class and ethnic profiles. Then males and females are subjected to the same intervention and any outcome differences are attributed to differences of gender. Sometimes the same effect is created through a matched pairs design (Bowling, 1997: 219). Sometimes factorialisation is done to create sub-groups within an RCT.

Reference group/area designs

There is an experimental group who are subjected to some intervention. The results of this are compared with the 'before and after' characteristics of some other group deemed to be similar but who have not received the intervention, or have received a different one. For example, the smoking behaviour of children in a school in which there has been anti-smoking health education is compared with the smoking behaviour in a school where there has been no such health education campaign. Attempts are made to select areas/groups who are similar in at least their demographic characteristics: age, gender, deprivation, ethnicity, for example. However, the possibilities for confounding factors to muddy the results are many. This is the other major strategy used in 'case control studies' (see Chapter 10).

Pre- and post-test designs without controls

Subjects for the experiment are studied prior to intervention to provide some baseline measures. All are subject to the intervention, the outcomes are measured and compared with the baseline measures. In effect 'the controls' are these same people prior to intervention. However, there is rarely any way of deciding which of the differences between the pre- and post-test were due to the intervention, which were due to the characteristics of the people chosen (for example due to their propensity for spontaneous remission from a condition being treated), or to the elapsing of time, or which were due to the fact that they were subjects of an experiment, over and above the effects of the intervention itself. Where designs like this are used to investigate causality, including effectiveness, little confidence should be placed in the results.

* In this box 'intervention' is used as a shorthand to include treatments such as drug treatments or surgical operations, exposure to health education, exposure to pathogens (as in influenza research), or modes of care delivery (such as care management) and so on.

people is actually a sample of one and that the experimental 'subject' is the group as a whole. It is a moot point as to whether the healing of an ulcer on one leg is independent of the healing of an ulcer on the other leg of the same person. The experiment in Chapter 1 involved 35 ulcerated legs and 29 people. Was this a sample of 35 or of 29?

5 Subject reactions, researcher bias and blinding

(See boxes 2, 3 and 5 on Figure 5.2)
The understandings of subjects, practitioners and researchers will, of course, influence what happens in an experiment. That cannot be avoided. What can be avoided is such influences affecting one arm of

the trial more or differently from another. Preventing subjects, practitioners and researchers knowing which arm of the trial particular subjects are in standardises this influence across all arms of the trial. This is rather unfortunately known as 'blinding'.

There is a large, and long-established, research literature showing how the expectancies of subjects can influence research outcomes, and how the unwitting biases of practitioners and researchers can influence the results. Although the physiological processes of wound healing might seem immune to this influence, there is research showing that the expectations of practitioners can affect the speed at which leg ulcers heal (Schwartz et al., 1988). The 'double-blind randomised controlled trial' is one where neither the subjects themselves nor the practitioners/researchers know to which arm of the experiment subjects belong, group membership only being disclosed at the end of the trial or earlier if it has to be broken off prematurely for safety reasons. Random allocation helps with blinding, but blinding usually also involves secret codes, sealed envelopes and perhaps the use of an independent agency to allocate subjects to treatments. The 'randomisation centre' in Figure 5.4 blindly allocated patients to treatments, though it did not blind patients or practitioners to the mode of care patients received. Blinding may also involve making treatments look, taste or feel similar; for example two different drugs or a drug and a placebo (a dummy drug), made up into tablets of identical appearance.

It is impossible or unethical to blind practitioners/researchers and difficult to blind subjects where the intervention is surgery, or diet, a health education programme, or counselling, and in the study in Chapter 3 it was impossible to blind either as to whether patients were being cared for in hospital or at home. Where blinding is impossible, confidence in the results must be lower because there is always a possibility that any outcome differences came from knowledge of what the treatment was, rather than from the effects of the treatment itself. Schultz and colleagues suggest that the absence of double-blinding may create spurious outcome differences of up to 17 per cent (Schultz et al., 1995), but other research suggests even greater effects (Rosenthal and Rubin, 1978).

6 Regression to the mean

(See box 7 on Figure 5.2)
Regression to the mean is a particularly common cause of confounding in experimental research. Imagine that you threw six dice and gained a score of 9 – you threw all 1s and 2s. You know intuitively that if you were to throw the dice a second time you would be likely to get a higher score. This is simply because there are more other scores to score than 1s and 2s. The same regression effect would be true if on

the first throw you scored mainly 5s and 6s. Then the next throw would be more likely to result in a lower score. Similarly, if subjects for an experiment are chosen because they have extreme scores for some illness, or behaviour or social problem, chance alone is likely to mean that they will show improvement over time. In investigations without controls there is no way of knowing whether all improvement shown is due to this statistical artefact, that is, changes that are not due to the intervention at all. If one arm of an experiment started with subjects with more extreme scores than the other, more improvement shown for that arm may be due to the regression effect. Unfortunately usual tests for statistical significance (see Chapter 7) will not distinguish regression effects from real effects. More complicated statistical analysis is needed to do this (Senn, 1997). However, even without statistical analysis readers of research may have a reasonable suspicion that the results are affected by regression to the mean if the subjects in one arm of the trial have notably more extreme baseline scores than the subjects in the other arm *and* if those with the more extreme scores also show the greatest improvement.

7 Replicability

As the philosopher Daniel Dennett says, science is 'making mistakes for all to see, in the hopes of getting others to help with the corrections' (1995: 380). For experiments this means researchers specifying exactly what was done, so that, in principle at least, someone else can repeat the experiment. The possibility of replication is what makes experimental researchers so much more accountable to their readers than other kinds of researchers. RCTs in medicine are one of the fields of science where replication, or near replication, is widely practised. As Chapter 4 shows, practitioners may often be able to draw on evidence from several experiments of much the same kind. In applying research results to practice it is also important that what was done in the experiment was clearly specified. If it is not, it will be impossible for practitioners to know how to 'do the same', and hence difficult for them to achieve similar results. Achieving replicability may be difficult with complex interventions – see section 9 and Chapter 2).

8 Intention to treat

(See box 8 on Figure 5.2)
It is not essential that each arm of an experiment has equal numbers of subjects. But once under way, it is problematic if people drop out. This is because differences in outcomes between arms might then be due to the people dropping out from one arm of the experiment being different from those dropping out from another arm. Drop out may, or

may not, have to do with the treatments received. Prejudices on behalf of practitioners may influence the drop out. Sometimes, for therapeutic reasons, a subject will be swapped from an arm of the trial which receives a placebo treatment, to an arm which receives an active treatment; an unplanned cross-over. Sometimes subjects do not comply with their treatment, and do something which is very like what happens in a different arm of the trial. The way in which drop outs, unplanned cross-overs and non-compliance should be managed is to regard all subjects as still belonging to the groups to which they were originally allocated. This is called reporting the results in terms of *intention to treat*. Thus in a diet experiment where someone on a low fat diet cheats and eats a high fat diet, s/he will still be counted as a member of the low fat arm of the trial. There may, of course, be problems of researchers knowing about compliance and non-compliance.

At the end of the trial the subjects remaining should be analysed to see whether the characteristics of the groups have changed due to drop out, cross-over and so on, and whether this might have influenced the results.

9 Simple and complex interventions

In the blinded RCT, treatments should seem similar to subjects and practitioners, thus eliminating one possibility for confounding. And whether blinded or not, most medical RCTs conducted in academia put a premium on a standardised performance by practitioners, using written protocols and sometimes special training, specifying what should be done, how measurements should be taken and how judgements should be made. However, there has been considerable concern expressed about laxity in these regards concerning RCTs by drug companies, particularly those carried out in general practice (*British Medical Journal*, 1998; Boseley, 1999).

Standardisation is important if the experiment is to be replicable (see section 7 above and box 9 on Figure 5.2). Standardisation is more possible where treatments are simple to administer and it is justifiable to administer them in a standardised way. This is often not the case. Counselling, psychotherapy, most social work interventions and some nursing interventions are customised to the particularities of each individual client. If this happens in experimental research, then each arm of the trial will actually contain not the same, but diverse treatments. There may then be more similarities between some treatments in different arms of the experiment, than between some treatments in the same arm of the experiment. In fact, the overwhelming majority of experimental studies in social work and counselling show 'no difference' between 'different treatments' (Newman, 1994; Oakley and Fullerton, 1996). This is sometimes interpreted benignly

through issuing the dodo bird verdict – 'everyone a winner', and sometimes adversely with the verdict that there is no evidence that social work or counselling are more effective than leaving people alone (Brugha and Glover, 1998). But it is equally likely to be due to the fact that these interventions are so variable in practice that, either they are not amenable to experimental research, or, if they are, they would need enormously large samples to accommodate the diversity *within* arms of the trial (Chapter 7, section 7) and a great effort to record the differences accurately.

Complex interventions then, make for difficulties in replication (section 7 above). Where interventions are complex it may be difficult for researchers to know which are the important ingredients of the intervention and hence which to record. Similarly, complex interventions will be more difficult for practitioners to emulate. The experiment which features in Chapter 2 exemplifies these problems.

10 Reliable and sensitive measurements

(See boxes 4, 5 and 10 on Figure 5.2)
It is crucial that the same assessments and measurements are made of all subjects, each in the same way, otherwise it will be unclear as to whether different baseline and outcome measures reflect real differences, or merely inconsistency in measurement. If practitioners are not blinded to the intervention subjects have received, it is common to involve an independent party who does not know this to do the assessments to avoid bias contaminating the measurements.

Chapter 6 discusses measurement instruments, but here it is worth noting that insensitive instruments may give results which are spuriously similar. For example, designating ulcers as merely 'healed' or 'not healed' may miss the fact that some 'not-healed' ulcers are more healed than others and perhaps making the superiority of one treatment over another invisible. Unfortunately, measurement instruments which discriminate finely also set up the ideal conditions for regression effects (section 6 above), since there are more positions on the scale for subjects to regress from. Consistency of judgement by practitioners and researchers is also important. Testing for this is dealt with in Chapter 6, section 6.

11 The internal validity of experiments

Roughly speaking, internal validity refers to whether what researchers claim to be true is indeed true for the subjects in the setting in which they did their research. Much of the critical appraisal of the internal validity of any experiment revolves around three kinds of question which appear in different versions in Figure 5.2 and feature in much

more detail in 'Questions to Ask about Experiments' in Part 4 of this book:

- Were the subjects really similar in the ways they should be similar and/or really different in the ways they should be different?
- Were the treatments really different in the ways they should be different and/or similar in the ways they should be similar?
- Were any other influences at play which should have been excluded, actually excluded from having an effect on the outcomes?

Only if all these questions can be answered in the affirmative can the results of an experiment be accepted with confidence. These same questions, or something like them, arise wherever someone makes a claim that doing such and such has such and such an outcome. In the absence of experimental control they are very difficult questions to answer convincingly.

12 The external validity of experiments

A research study may be valid in its own terms – internally valid – but none the less what happened in the experiment might not happen anywhere or anytime else; its results may not be generalisable; they may lack external validity.

The way in which experiments control variables is what makes them good designs for studying causality. Unfortunately in the real world variables may not be controlled, or controllable. Thus the artificiality of the experiment can become a problem when attempts are made to generalise from the experiment to situations beyond it. The success of physics and chemistry in producing knowledge with everyday applications does not come entirely from using the experimental method, but also from transforming the world so that what happens in the laboratory can be made to happen outside it. High tech machine medicine works because it makes hospitals more like experimental laboratories where variables can indeed be brought under greater control. Thus when it comes to making a judgement about the generalisability of research findings the question is whether enough of the circumstances under which the experiment was performed could be reproduced in practice to ensure that similar results are obtained. Some kinds of interventions, drug treatments for example, do often seem to travel quite well, whereas the outcomes of others seem to depend very much on who does them, where, to whom and under what circumstances. Table 5.1 suggests the characteristics of topics with regard to which research might produce more or less generalisable knowledge about effectiveness.

Another issue about the generalisability of experimental research concerns the representativeness of the people involved as subjects. In

Table 5.1 Topics for which research is more or less likely to produce sound generalisations

The characteristics of topics about which research is more likely to produce generalisable knowledge about effectiveness	The characteristics of topics about which research is less likely to produce generalisable knowledge about effectiveness
Where the entities being studied have robust, reliable and predictable properties: for example, materials, forces, muscles, bones, cells, genes	Where the entities being studied do not have robust, reliable and predictable properties; for example, emotions, interpretations, meanings, relationships, group dynamics
Where interventions are simple and standardised and can be much the same irrespective of which practitioner carries them out; for example, administering a drug	Where interventions are complex and differ from client to client and/or where the same named intervention differs according to which practitioner carries it out; for example, counselling
Where there is a strong consensus (or an enforcement) of some criteria of effectiveness	Where there are multiple and contradictory criteria of effectiveness
Where there is little ambiguity as to what evidence counts as meeting the criteria of effectiveness	Where there is considerable ambiguity as to what evidence counts as meeting criteria of effectiveness

fact, it is relatively rare for RCTs in health and social care to start with representative samples. They usually use 'grab' or 'convenience' samples: just the people who happen to be around when subjects are needed. Or they may be samples selected so that any effect of different treatments will show clearly in the results. For example, Tudor Hart (1993) notes that most trials for hypertension management involve people who are suffering from nothing other than hypertension. But 90 per cent of people suffering from hypertension are suffering from something else as well. Selecting those with hypertension alone gives the best chance of achieving clear unambiguous results not muddled by other conditions from which subjects might be suffering. But since such people are unlike most of those with hypertension who crop up in routine practice, it will be difficult for practitioners to know how far what worked for people suffering only from hypertension will work for people suffering from a diversity of other conditions as well – people like their own patients.

Practitioners do not recruit their own clients by representative sampling either, so for most practical purposes it is not too important that an experiment has subjects who are unrepresentative of some wider population. What is important is that researchers publish enough details about the subjects so that practitioners can extrapolate the results to their own distinctive client mix. Statements about indications and contraindications are particularly useful in this regard.

Differences of client mix and of practice circumstances together may mean that what 'worked' in the research will not work in practice.

Research in the cost-effectiveness field is particularly sensitive to context, because doing something in one place very rarely costs the same as doing it in another, to say nothing of the problem that the cost-base of practice is likely to change quite quickly through time (see also Chapter 7, section 12). Thus economic analyses have a two-way problem of generalisability. The relative effectiveness of two treatments shown under research conditions may not be the same as can be achieved in some practice setting, *and* the relative costs of two interventions demonstrated in the research may not be the same as happens in practice, even in the same practice after a period of time.

A further problem with regard to generalisation comes from the interpretations made by subjects and practitioners. Blinding can prevent such interpretations affecting outcomes for one arm of the trial more than another. But it cannot prevent such interpretations influencing the outcomes for (at least some) subjects in *all* arms of the trial. In routine practice clients are not blinded to the treatment they receive, and this makes it possible that some effects that occur because people are blinded, or some that occur because people know they are part of an experiment, will not be reproduced in practice. Or again, under experimental conditions it may be possible for practitioners to ensure a high degree of compliance with a treatment – much higher than would be possible in routine practice. Similarly, staff may be particularly punctilious, or enthusiastic or otherwise different in an experiment compared with their counterparts in routine practice. The term *Hawthorne effect* is often used to refer to the tendency for people to behave differently because they know they are being researched or involved in research (Sapsford and Abbott, 1992: 105), while the term *experiment effect* may be used to include this, and any other things that are more likely to happen in an experiment than under naturally occurring circumstances. This is one reason for the sensitivity analysis carried out by Shepperd et al., in the exemplar study in Chapter 3 (see also Chapter 7, section 12).

13 Single subject experiments

The fact that experiments typically produce results for groups of people often makes it difficult for practitioners to apply the results to individual clients, because what is 'effective' for a group may be beneficial to some of them, have no effects on others, and adverse effects on yet others. The techniques of n-of-1 experiments in medicine and single case evaluations in social work and clinical psychology take the experimental approach down to the individual level.

The 'n' in the n-of-1 refers to the number of subjects in the experiment, which is one person. But sometimes several are conducted

in parallel with different patients, making the term n-of-1 slightly misleading.

The more usual kinds of randomised controlled trial provide predictions about what is likely to happen among a group of people subject to a treatment, but do not provide enough information to predict what will happen to any patient in particular. In this sense n-of-1s bridge the gap between knowledge about groups and knowledge about individuals. One way of looking at n-of-1 trials is as their being a more systematic version of the kinds trial and error procedures used by clinicians when, for instance, they attempt to stabilise a drug dose for a patient, or to establish the cause of a food allergy by systematically excluding items of diet. N-of-1 trials are usually based on what has been found effective *on average* for groups, where what is unknown is the effectiveness of the treatment for the individual patient.

Box 5.2 provides an example of a typical n-of-1 trial, or rather of a set of n-of-1s being conducted alongside each other.

These are double-blind randomised controlled experiments with cross-over (see Box 5.1). However, it is not patients who are randomly allocated to different treatments but different treatments which are allocated to the *same* patient in a random sequence (two different treatments, or a treatment and a placebo). Where there is indeed only one trial going on, the only purpose of randomisation is the double-blinding (see section 5 above). With a series of n-of-1 experiments, randomisation will also control for the effects of the order in which treatments are given. Such control of order-effects will apply only to the results of all the trials, and not to the result of any one of them.

This is a 'same-subject(s) cross-over' design where the individual patient constitutes his or her own 'control' in the sense that what happens to the patient when receiving one treatment, is compared with what happens to the same patient when receiving the alternative treatment (the cross-over) (see Figure 5.5 in Box 5.2). This limits the use of n-of-1 experiments to the kinds of interventions which have only transitory effects. Pain control is a particularly common area. As always, the purpose of an experimental approach is to control variables that might influence outcomes, other than the intervention which is given an opportunity to effect outcomes. In any RCT there is always a possibility that other things happening in a patient's life will affect their response to treatment. Where there are many patients in the trial, as in the usual RCT, then there is a good chance that these extraneous variables will balance out between the arms of the trial. But in an n-of-1 trial there is only one patient who has, in random sequence, to be in both arms of the trial. Since the primary purpose of the trial is to establish what works for this particular patient the clinician will want to discover the treatment which works best, extraneous factors and all. Extraneous factors that are persistent in the patient's life, or show a consistent time trend do not constitute a

research problem, but extraneous factors that come and go may well do so. Thus if the administration of the active drug coincided with a particularly stressful period at work, and the placebo with the patient's holidays, this may well give misleading results. Hence a run of several alternations (or cycles) of treatments is important, despite the strong temptation to discontinue the trial when the patient feels better, or feels particularly ill. Drop out is particularly common in n-of-1s. Group RCTs can survive drop outs, but drop out from an n-of-1 experiment is the end of the affair.

Series of n-of-1 trials demonstrate the diversity of responses to treatments. They may provide useful data about indications and contraindications which can be used by practitioners in their own decision-making. With the pressure towards 'evidence based practice' some clinicians have seen the n-of-1 experiment as something which should be adopted in routine practice where possible, as a way of customising to the individual the knowledge derived from effectiveness research on groups (Campbell, 1994).

The human genome project has, on the one hand, thrown some doubt on the standard RCT in medical research, and on the other has directed interest towards n-of-1 experiments as an alternative. The source of the doubt is the way that the human genome project draws attention to the genetic diversity which must exist among any group of subjects for an RCT and hence to the possibility of a large field of uncontrolled variables. The larger the diversity, the larger a

Box 5.2 N-of-1 trials comparing a non-steroidal anti-inflammatory drug (NSAID) with paracetamol for osteoarthritis (based on March et al., 1994)

Paracetamol has a risk of accidental or intended overdose. NSAID drugs carry a risk of gastrointestinal complications, occasionally fatal. Thus it is important to choose the drug which, for the particular patient, gives the lowest risk for the optimum pain control. Where equally effective, paracetamol is usually to be preferred as the drug which carries the lower risk for most patients.

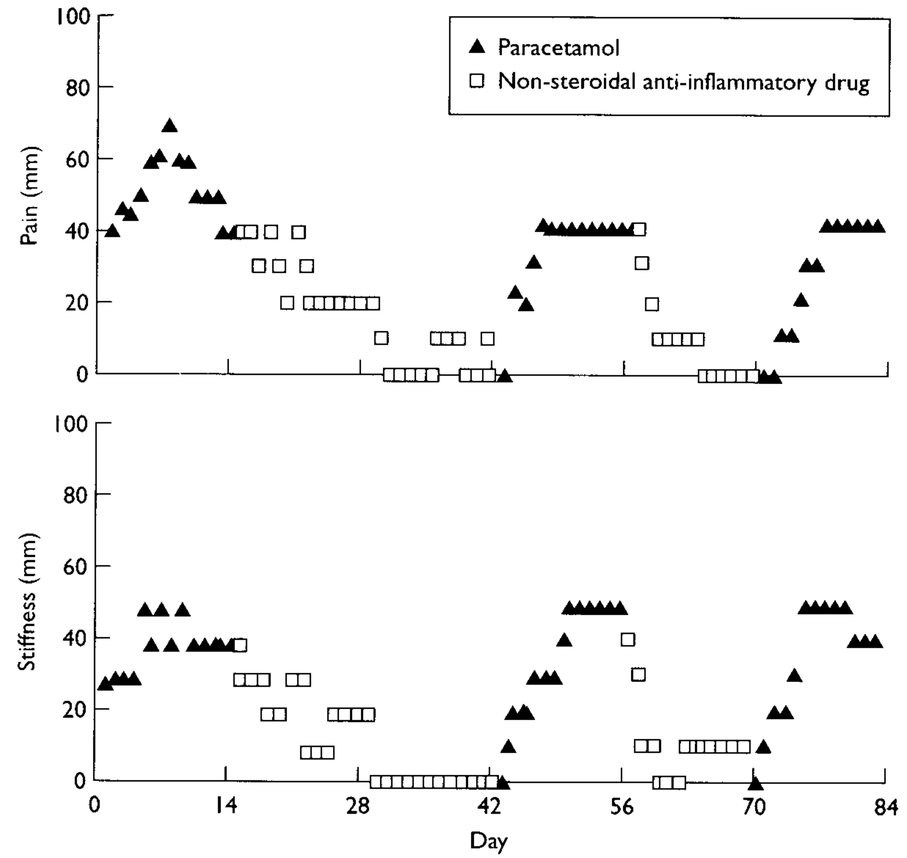
Main objective: to evaluate individual patient responses to paracetamol and a non-steroidal anti-inflammatory drug (NSAID) in terms of pain relief, immediate side effects and general well-being.

Subjects: 25 patients from general practice experiencing painful osteoarthritis with no contraindications for non-steroidal anti-inflammatories and no corticosteroid injection in the previous four weeks. Five patients dropped out very early, five dropped out before the end, but after a clinical decision had been made as to the best treatment for them.

Procedure: each patient was treated in three cycles of four weeks. Each cycle consisted of two weeks taking a non-steroidal anti-inflammatory drug and two weeks taking paracetamol. The order of this was determined for each patient at random. Both drugs appeared visually identical to the patient and researcher/practitioner and both were blind to which drug was in use. Patients were allowed to use paracetamol as 'escape' analgesic, whichever other treatment they were experiencing.

Measures: Patients completed a daily diary for pain and stiffness using visual analogue scales (see Figure 5.5) and a weekly checklist of 11 symptoms known to be associated with NSAID drugs. There was weekly monitoring of ability to perform activities selected according to which joints were inflamed. The use of escape analgesia was recorded. The most important outcome was the evidence necessary to make an informed clinical decision of the drug of choice acceptable to the patient.

Figure 5.5 Daily visual analogue scores for pain and stiffness for a patient who benefited most from the NSAID drug (March et al., 1994: Fig. 3: 1043)



Results: Fifteen patients completed the 12 weeks. For nine of these, paracetamol proved at least as, or more effective in pain control than NSAID. For the remainder, either NSAID proved the drug of choice without evidence of adverse side effects or both drugs were equally ineffectual or unacceptable.

sample needs to be to allow randomisation to produce two groups of subjects similar to each other in ways relevant to the experiment. Insofar as genetic differences do make a difference to the results of medical interventions, it is likely that the results of RCTs will reflect pre-existing genetic differences between subjects which have been inadequately randomised between the arms of an experiment if the sample has been too small. As the mapping of the human genome proceeds, it will become more and more possible to see what genetic differences do exist between people and to attempt to relate these differences to different responses to treatments. Knowing an individual's genome will also make it possible to control for genetic differences by starting with a pool of subjects selected for their genetic similarity in ways relevant to the experiment, and then randomising these between the arms of the experiment. This is a long established procedure in RCTs in agriculture and in medical research using animal subjects. In both fields experimenters usually start with a pool of subjects bred to be genetically very similar to each other. N-of-1 experiments offer a different way of controlling for genetic diversity since the same person is alternately 'experimental subject' and 'control'.

As with an n-of-1 trial in medicine, a single case evaluation is a same-subject(s) cross-over design and a way of discovering whether an intervention is effective for a particular client. Randomisation in the n-of-1 design is largely for the purpose of blinding. Thus it is not useful where it is neither feasible nor desirable to prevent clients and practitioners knowing which treatments are being administered. What is left of an n-of-1 structure when randomisation is removed is the alternation of interventions, or the alternation of a period of intervention with a period of non-intervention. That essentially is what a single case evaluation is. The technique has mainly been used in conjunction with cognitive behavioural interventions in social work and clinical psychology (Bloom and Fischer, 1982; Sheldon, 1983; Barlow and Hersen, 1984; Johannessen, 1991; Thyer, 1993; Kazi and Wilson, 1996). Box 5.3 gives an example.

Figure 5.6 in Box 5.3 shows that AR's attendance improves at the beginning of an intervention period, and then drops off, and that it deteriorates markedly when the intervention is withdrawn. During each intervention phase his attendance is on average higher than in the preceding phase of non-intervention. Overall there is a ratchet effect such that as time goes on his attendance during the period of non-intervention improves.

It does seem as if progress has been made with AR's attendance. That, of course, might have been due to factors other than the intervention made. However, alternating intervention and non-intervention and carefully recording the results provides evidence which suggests

Box 5.3 An example of single case evaluation in social work: an ABABA design

Figure 5.6 AR's school attendance (percentage of possible attendance per week) (based on Kazi and Wilson, 1996: Fig. 7: 707)

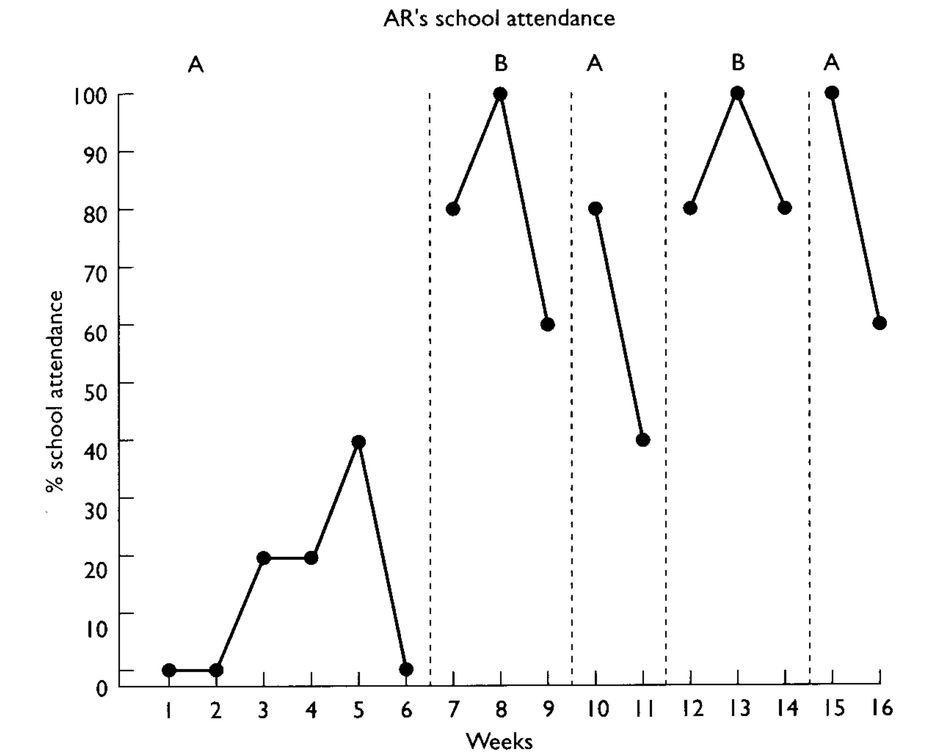


Figure 5.6 above records alternate phases of non-intervention (conventionally called 'A') and intervention (conventionally called 'B') with regard to AR, a 14-year-old referred to educational social work for persistent absenteeism by a Scottish High school. The interventions (Bs) consisted of counselling for both AR and his family, building better home–school links, and providing encouragement for AR in school. The measure used was the school register which was completed twice a day.

that the improvement is attributable to the intervention. Unless there were evidence that something other than the intervention caused the improvement in attendance, common-sense suggests that the evidence in Figure 5.6 should be accepted as evidence of effectiveness. This is particularly so if similar patterns of response are shown for the same procedures applied to other truants.

The intervention in the example was a complex one, and it is difficult to know which aspect of it was the effective ingredient; family

or individual counselling, better home–school links, more encouragement for AR in school, or indeed AR's perceptions that his absence might be more easily spotted. And each of these is again made up of many components. To investigate this further it is imaginable that a whole series of single case evaluations featuring AR might be mounted, each featuring only one aspect of the intervention. However, this does not seem desirable or feasible. First, the main purpose of single case evaluation is to solve the problem. Explaining how the trick was done is only a secondary objective. Second, while the outcome measure chosen here was school attendance, absenteeism was not the only problem being addressed. No doubt counselling, better home–school liaison and so on, are also addressing other issues, and in ways that would not be closely reflected in the school attendance record. Third, it is to be hoped that the problem would have been resolved in a period much shorter than would be required to test the intervention bit by bit.

Single case evaluation not only lends itself well to incorporation into routine practice but is likely to improve practice by requiring practitioners to specify targets carefully and to look for ways of measuring their attainment. Measures need to be valid and reliable (see Chapter 6), but since this is as much practice as research, measurement data also need to be of the kind which can be acquired with minimum effort by practitioners. It is usually beneficial if they are meaningful to the subject of the evaluation. In this example AR had the goals and the measurements foisted on him. But there is no reason why this approach should not be directed towards achieving goals chosen by the client and measured by indicators meaningful to him or her. This makes the approach particularly applicable to various kinds of 'brief intervention' counselling, or methods using contracts in social work or probation work.

There is a potential problem of 'false measurement'. In the example in Box 5.3, for instance, an improvement in school attendance might be accompanied by an increase in AR's misery. But given the nature of the intervention, this seems unlikely to go unnoticed.

14 Questions to ask about controlled experiments

In Part 4 of this book there is a checklist of questions to use in critically appraising pieces of experimental research. The questions step through the issues raised in this chapter, and some which are raised in Chapters 6 and 7. There are two kinds of questions: one kind is concerned with internal validity – does the research seem true in its own terms? The other kind relates to external validity or generalisability: is it likely that what happened in the experiment could be made to happen elsewhere, if so where, and would this be desirable?

You might like to use the checklist as a way of appraising the research reported in Chapters 1 or 2, or for appraising other published experimental research more closely related to your own interests.

15 Further reading on controlled experiments in health and social care research and cost-effectiveness studies

A good place to start is with Ann Bowling's *Research Methods in Health* (1997), or the chapter on RCTs by Shepperd et al. (1997), both of which also provide access to a wide range of more technical literature. Pocock's *Clinical Trials* (1983) is a standard reference for medical research. Campbell and Stanley (1996) provide a more than usually adequate account of quasi-experimental designs. Since the 1950s social work has been rather hostile to experimental research, though pre-war experimentalism was more common in social work than in medicine. Something of the contemporary debate is captured in Oakley and Fullerton (1996) and in the papers in Williams et al. (1999).

For single subject experiments, the original sources from which the examples in the chapter were drawn are themselves worth reading in their entirety (March et al., 1994; Kazi and Wilson, 1996). In addition, Johannessen (1991) is a useful source on n-of-1 trials; and see Sheldon (1983) and Thyer (1993) on single case evaluations.

Cost-effectiveness studies

Experimental designs are often used as a basis for judging the cost-effectiveness of interventions. Watson (1997) provides an introduction to this kind of analysis. Jefferson et al. (1996) is a more comprehensive guide. There are references on costing services at the end of Chapter 7.

References and further reading

Barlow, D. and Hersen, M. (1984) *Single Case Experimental Designs: Strategies for Studying Behaviour Change*, 2nd edition. London: Pergamon Press.

Bloom, M. and Fischer. J. (1982) *Evaluating Practice: Guidelines for the Accountable Professional*. Englewood Cliffs, NJ: Prentice Hall.

Boseley, S. (1999) 'Trial and error puts patients at risk', *Guardian*, Tuesday 27th July, p. 8.

Bowling, A. (1997) *Research Methods in Health: Investigating Health and Health Services*. Buckingham: Open University Press.

British Medical Journal (1998) 'Dealing with Research Misconduct in the United Kingdom', *British Medical Journal*, 316: 1726–33.

Brugha, T. and Glover, G. (1998) 'Process and health outcomes: need for clarity in systematic reviews of case management for severe mental disorders', *Health Trends*, 30 (3), 76–9.
Campbell, M. (1994) 'Commentary: n-of-1 trials may be useful for informed decision making', *British Medical Journal*, 309: 1044–5.
Campbell, T. and Stanley, J. (1996) *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally.
Dennett, D. (1995) *Darwin's Dangerous Idea: Evolution and the Meanings of Life*. London: Allen Lane.
Jefferson, T., Demicheli, V. and Mugford, M. (1996) *Elementary Economic Evaluation in Health Care*. London: BMJ Publications.
Johannessen, T. (1991) 'Controlled trials in single subjects', *British Medical Journal*, 303: 173–4.
Kazi, M. and Wilson, J. (1996) 'Applying single-case evaluation in social work', *British Journal of Social Work*, 26: 699–717.
March, L., Irwig, L., Schwartz, J., Simpson, J., Chock, C. and Brooks, P. (1994) 'n of 1 trials comparing a non-steroidal anti-inflammatory drug with paracetamol in osteoarthritis', *British Medical Journal*, 309: 1041–6.
Newman, F. (1994) 'Disabuse of the drug metaphor: introduction', *Journal of Consulting and Clinical Psychology*, 62 (5): 941.
Oakley, A. and Fullerton, D. (1996) 'The lamp-post of research: support or illumination?', in A. Oakley and H. Roberts (eds), *Evaluating Social Interventions*. Essex: Barnados, pp. 4–38.
Pocock, S. (1983) *Clinical Trials: a Practical Approach*. New York: John Wiley.
Roberts, C. and Sibbald, B. (1998) 'Randomising groups of patients', *British Medical Journal*, 316: 1898.
Rosenthal, R. and Rubin, D. (1978) 'Interpersonal expectancy effects: the first 345 studies', *The Behavioural and Brain Sciences*, 3: 377–415.
Sapsford, R. and Abbott, P. (1992) *Research Methods for Nurses and the Caring Professions*. Buckingham: Open University Press.
Schultz, K., Charmers, I., Hayes, R. and Altman, D. (1995) 'Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials', *Journal of American Medical Association*, 273: 408–12.
Schwartz, S., Flamant, R. and Lellouch, J. (1988) *Clinical Trials* (trans. M. Healy). London: Academic Press.
Senn, S. (1997) 'Regression to the mean', *Statistical Methods in Medical Research*, 6: 99–102.
Sheldon, B. (1983) 'The use of single case experimental designs in the evaluation of social work', *British Journal of Social Work*, 13: 477–500.
Shepperd, S., Doll, H. and Jenkinson, C. (1997) 'Randomized controlled trials', in C. Jenkinson. (ed.), *Assessment and Evaluation of Health and Medical Care: a Methods Text*. Buckingham: Open University Press. pp. 6–30.
Shepperd, S., Harwood, D., Jenkinson, C., Gray, A., Vessey, M. and Morgan, P. (1998) 'Randomised controlled trial comparing hospital at home care with inpatient hospital care. I: three month follow up of health outcomes', *British Medical Journal*, 316: 1786–91.
Thyer, B. (1993) 'Single-system research designs', in R. Grinnerll (ed.), *Social Work Research and Evaluation*, 4th edition. Itasca, Ill., F.E. Peacock, pp. 94–117.
Tudor Hart, J. (1993) 'Hypertension guidelines: other diseases complicate management', *British Medical Journal*, 306: 1337.
Watson, K. (1997) 'Economic evaluation of health care', in C. Jenkinson (ed.), *Assessment and Evaluation of Health and Medical Care: a Methods Text*. Buckingham: Open University Press. pp. 129–50.
Williams, F., Popay, J. and Oakley, A. (1999) *Welfare Research: a Critical Review*. London: UCL Press.

CHAPTER 6

RESEARCH INSTRUMENTS IN EXPERIMENTAL RESEARCH

Introduction — 1 The use of standard instruments — 2 Cultural specificity and instruments — 3 Data levels and qualities — 4 Instruments and data distributions, data transformations, floor and ceiling effects — 5 Validating instruments for their reliability — 6 Reliability tests — 7 Validating the validity of instruments — 8 Questions to ask about research instruments — 9 Further reading on research instruments — References and further reading

Introduction

Experiments usually entail the use of some kind of data collection instrument to produce quantitative data which are then analysed statistically. The instrument used, the measurements made and the way the data are analysed shape the results. In appraising research it is important to know how the instruments and statistical tests shaped the results.

Any device that is used to aid data collection can be called an 'instrument' in research, ranging from thermometers and their associated temperature charts to questionnaires used in surveys. For research purposes, 'having a fever' may be scoring above a certain level on a thermometer, and 'being satisfied with the NHS' may be ticking a particular box on a questionnaire (see Chapter 8). Any instrument structures the data it collects. Thus looking at the design and use of such instruments is an extremely important aspect of appraising research.

Figure 6.1 gives an example of a research instrument: it shows one of the panels of the Dartmouth COOP charts used in the research by Shepperd and colleagues which is presented as the exemplar study in Chapter 3 of this volume. Table 3.1 in Chapter 3 gives data resulting from the use of the whole set of COOP charts.

Figure 6.1 gives a simple illustration of what a measurement instrument of this kind does, which is to turn ideas into numbers which can then be mathematically manipulated. It also hints at the kinds of queries this manoeuvre gives rise to. For example, what, exactly, is being measured? Nominally this is 'feelings over the past