4th Edition

Research Methods in Psychology

Edited by

Glynis M. Breakwell, Jonathan A. Smith and Daniel B. Wright

SAGE

Los Angeles | London | New Delhi
Singapore | Washington DC

SAGE

Los Angeles | London | New Delhi
Singapore | Washington DC

SAGE Publications Ltd
1 Oliver's Yard
55 City Road
London EC1Y 1SP

SAGE Publications Inc.
2455 Teller Road
Thousand Oaks, California 91320

SAGE Publications India Pvt Ltd
B 1/I 1 Mohan Cooperative Industrial Area
Mathura RoadA
New Delhi 110 044

SAGE Publications Asia-Pacific Pte Ltd
3 Church Street
#10-04 Samsung Hub
Singapore 049483

Editor: Michael Carmichael
Development editor: Robin Lupton
Editorial assistant: Alana Clogan
Production editor: Katherine Haw
Copyeditor: Cenveo Publisher Services
Proofreader: Cenveo Publisher Services
Marketing manager: Michael Ainsley
Cover design: Wendy Scott
Typeset by: Cenveo Publisher Services
Printed by: Ashford Colour Press Ltd,
Gosport, Hants

FSC
www.fsc.org
MIX
Paper from
responsible sources
FSC® C011748

First published 2012

Library of Congress Control Number: 2012930912

British Library Cataloguing in Publication data

A catalogue record for this book is available from the British Library

ISBN 978-0-85702-263-9
ISBN 978-0-85702-264-6 (pbk)

4 Quasi-experimental Designs

Chris Fife-Schaw

CONTENTS

AIMS OF THIS CHAPTER

This chapter deals with experiments where, for a variety of reasons, you do not have full control over the allocation of participants to experimental conditions as is required in true experiments. Three common quasi-experimental designs are described; the non-equivalent control group design, the time series design and the time series with non-equivalent control group design.

KEY TERMS

compensatory rivalry
contamination effects
external validity
history effects
instrumentation effects
internal validity
interrupted time series

maturational effects
multiple time series design
regression to the mean
sample selection bias
selection/ maturation interaction
subject or participant mortality
testing effects

4.1 INTRODUCTION

In Chapter 3 the basics of classical experimental designs were discussed. The value of doing experiments is that they offer the most clear-cut route to testing hypotheses about causes and effects. The experimenter has control over the relevant independent variables and allocates participants to conditions at random in an attempt to make sure that they know exactly what is responsible for the changes they observe.

This is to be contrasted with observational and correlational approaches, where we might be able to show that two variables appear to be related to one another but it is difficult to determine whether there is a causal relationship between the variables (where one 'causes' the other) or some third variable is responsible for the observed relationship. Although this may seem less than satisfactory – after all, we usually want to be able to say what causes what – correlational studies are often the best we can hope for in many real-world situations. Practical considerations may limit the amount of control we can expect to have in such situations, so we have to be careful whenever we try to interpret relationships between variables.

In between correlational and experimental approaches lie two other kinds of study: the pre-experiment and the quasi-experiment. Pre-experiments are best thought of as studies that are done simply to get an initial feel for what is going on in a particular situation prior to conducting a more rigorous investigation; this is probably best illustrated by an example.

4.2 PRE-EXPERIMENTS

I once attended a rapid-reading course in an attempt to increase the speed with which I could get through paperwork. The university was happy to supply this kind of training as it would help the staff perform better and this should, in turn, help the university to be more efficient. A consultant was hired to do the training. In line with the current political concern to evaluate everything, the consultant felt obliged to conduct an experiment to see if the training had actually worked. Before the training started we were given a report to read and we were asked to time our reading of it and answer some factual questions about the report's content. Having done this, the training went ahead and at the end of the day we were tested on our reading speed again. So that the times and test scores could be readily compared we read the same text and answered the same questions as before. Needless to say reading speed had increased dramatically (four times quicker in my case) and accuracy was very high. The consultant, with evident satisfaction, declared the day a success. Of course the problem here is that we do not really know if the training had any effect on reading speed at all. Whether we have been able to accurately detect the effect of the training is referred to as the internal validity of the experiment.

There are several problems with this evaluation which challenge its internal validity even though at first sight it looks like a reasonable thing to have done. First, the test materials were the same on both occasions and since we had seen them only about seven hours previously there is a strong possibility that we were remembering the content rather than displaying skills learned in the interim. Thus the improvements may have been reflecting memory for the material rather than any real increased reading speed. You do not need to be a psychologist to know that it is easier to read something quickly if you already know what it is about. The same applies to the 'test' questions. Such threats to the experiment's internal validity are called testing effects. In all sorts of studies, repeatedly exposing participants to the test materials is likely to make them familiar with them and less anxious about what they have to do. Such effects tend to inflate post-test scores. In fairness, were the consultant to have used a different report and different test questions, it would have been even more difficult to know what any differences in reading speed could be attributed to. The second text might have been naturally easier to read or, possibly, more difficult.

A second problem concerns what are called maturational effects. Merely having the time to concentrate on reading speed even without experiencing the training may have led to improvements. As none of those tested had been allowed to spend the day thinking about rapid reading without also being exposed to the training, we do not really know whether the training itself had an effect.

Another problem concerns sample selection bias. All those present felt that they had a reading speed problem and, at least at the start of the day, were motivated to improve. You had to volunteer for the course and there was no external pressure on people to attend. Having put a day aside to improve reading skills, not trying hard to improve would have been somewhat perverse. This factor, in conjunction with the potential maturational effects noted above, may have served to increase scores on the retest. Again, we cannot really say how effective the training was, and even if it was effective here, it might be somewhat less useful when people are not so keen to be trained. This latter point refers to the external validity of the study: just how generalisable are the findings? If training works, does it only work for very committed people?

It should be noted that all of these problems concern the experiment (as a pre-experiment) and do not say anything about the virtues of the course. It may have worked very well or it may not. Whichever is the case, this study shed very little light on the issue. This is obviously not an ideal way to demonstrate that the training package increased reading speed.

Other common forms of pre-experiment are often found in news stories where some sort of intervention has to be evaluated. An example would be to see if peer teaching improved computing skills by comparing children's exam performances in schools that had adopted peer teaching with ones that had maintained traditional teacher-led methods. At one level this looks like a reasonable comparison between treatment groups – one group that gets peer teaching and one that does not. Clearly a true controlled experiment is not possible as it would be ethically and politically unacceptable to randomly allocate children to schools and thus to the 'treatment' conditions.

There are numerous problems in interpreting any differences that are observed between the groups. First, there is the question of whether the schools are comparable. Perhaps the schools that adopt peer teaching simply have more able or more socially advantaged children in them in the first place. Those children from better off backgrounds may be expected to have newer and better computers at home and be more computer literate, for instance. There is also a possibility that some event, such as a cutback in funds for computer maintenance, may occur in one school and not in another. Such a sudden change in one of the groups is known as a history effect and may lead to a difference between the groups which is not attributable to the treatment (here peer teaching) but is due to something else. While pre-experiments may seem so flawed as to be pointless, they do serve a purpose of highlighting problems that need to be addressed when the resources become available to do something more impressive and rigorous.

4.3 QUASI-EXPERIMENTS

Many of the problems discussed in relation to pre-experiments reduce the degree of certainty you can have that the 'treatment' actually caused the observed differences in the dependent variable of interest (i.e. the study's internal validity). Because of this, it is rare to see pre-experiments in academic journals. However, many of the research questions that we would like to answer simply cannot be answered by resorting to true experiments. This is usually because either we cannot randomly allocate participants to treatment conditions for practical reasons or it would be unethical to do so (e.g. if it would mean withholding treatment from someone who needs it). In the computer skills example above, for instance, we could not randomly allocate children to the schools.

Quasi-experiments should not be seen, however, as always inferior to true experiments. Sometimes quasi-experiments are the next logical step in a long research process where laboratory-based experimental findings need to be tested in practical situations to see if the findings are really useful. Laboratory-based experiments often reveal intriguing insights, yet the practical importance, or substantive significance, of these can only be assessed quasi-experimentally. Laboratory studies may have shown that under certain highly controlled conditions, peer teaching improves computer test scores, but the 'real' issue is whether peer teaching is generally a good thing for children in their schools. This is a question about the external validity of the laboratory-based studies.

Three classical quasi-experimental designs exist which attempt to overcome the threats to internal validity discussed above. What is presented below is a summary of the three prototypical designs; many variations of these are possible (see Cook & Campbell, 1979).

4.4 NON-EQUIVALENT CONTROL GROUP DESIGNS

As we saw in the example of the computer skills, the two groups (as defined by which school they attended) may not have been comparable. The intervention of peer teaching (the treatment) may have had an effect on test scores but we cannot be sure that the peer teaching group was not already better at computing, prior to the inception of the new programme. The non-equivalent control group design (NECG) overcomes this by requiring a pre-test of computing skill as well as a post-test. The pre-test allows us to have some idea of how similar the control and treatment group were before the intervention.

Figure 4.1 shows some possible outcomes from a simple NECG design. In graph A the control group starts off scoring less than the treatment group, reflecting the non-equivalence of the two groups; finding a control group with exactly equivalent scores in a quasi-experimental design is difficult. Both groups improve after the intervention but the treatment group has clearly improved more than the control group. This is quite a realistic picture to find in studies of educational interventions like the computer skills study. We would expect the control group to improve a bit as, after all, they are still being taught and are still maturing. If the treatment had an effect, then scores should have improved more than might have been expected if the intervention had not taken place. Graph B shows what might have happened if the treatment had no effect. Scores in both groups change about the same amount.

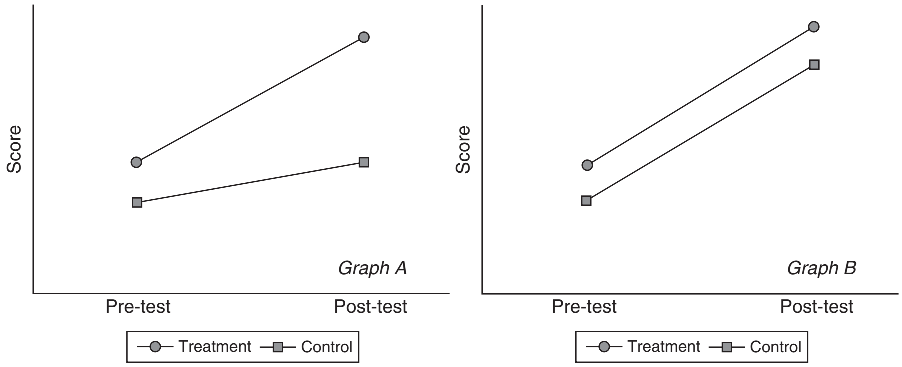

Figure 4.1 Non-equivalent control group designs

The graphs in Figure 4.1 are prototypical and reflect improvements over time. It is, of course, possible for all sorts of patterns to be found. Non-equivalent controls may outscore the treatment group at the pre-test; they may even be equal. Perhaps a treatment serves to allow the treatment group to 'catch up' with the controls. The treatment might decrease scores. There are many possibilities. In all cases you are looking for an interaction between treatment condition (treatment vs. control) and time of measurement (pre-test vs. post-test). You would obviously test for such an interaction statistically (see Chapter 10), but by plotting graphs like these you should observe lines of differing gradients; parallel lines usually indicate no treatment effect (but see later).

4.4.1 Problems with NECG designs

Almost by definition, NECG designs suffer from potential sample selection biases. In studies of 'alternative' therapeutic interventions in particular, there is often a problem that those who get a new treatment had actually sought it out, perhaps because traditional treatments had not worked for them. Such people may be highly motivated to see the new treatment succeed and might have ideological objections to existing treatments. There is also the possibility that those offering the therapy may, consciously or unconsciously, select people they believe would benefit from it or who they think will comply with the treatment regimen. Those who are thought likely to be 'difficult' cases, or for whom the disease may have progressed too far, might not be selected and may even end up appearing in the control group.

Clearly it would be unethical to refuse a new treatment to those who want it or to force those content with existing treatments to receive a new but still untested treatment. However, where possible, you should attempt to have control over, or at least full knowledge of, how the samples are selected. Be aware that those whose efforts are being evaluated will have a vested interest in the outcome of your study.

Even though we have pre-test measures by which we can compare samples, this does not guarantee that the two groups were truly equivalent before the treatment started. If one group was more able or 'brighter', maturation may proceed at a faster rate in that group than the other. We might expect, for instance, that children's computer skills improve with age (maturation) and that more able children learn these skills more quickly and easily. Were the treatment group to contain proportionately more high ability children, group differences may arise out of these differential rates of maturation rather than exposure to the peer teaching method. This is referred to as a selection–maturation interaction. As the pre-test is usually only used to compare groups on the dependent variable, such a problem may remain undetected. One obvious solution would be to measure variables that might conceivably lead to differential maturation rates at the pre-test (e.g. IQ), though this also increases demands on participants.

Statistical regression to the mean is another phenomenon which may influence interpretation of the data. Regression towards the mean is reflected in very high pre-test scorers scoring lower at post-test and very low pre-test scorers scoring higher at post-test. If we are studying people who score at the extremes on the dependent variable we may mistake changes at post-test for this regression to the mean. Why this happens is a little difficult to grasp at first but depends on the fact that our test measures will inevitably contain some errors (see Chapter 7). Cook and Campbell (1979) use an everyday example which is fairly easy to understand; the following is an embellished version of their example.

If we have an ability test like an exam we might do worse than our 'true' ability because we were distracted by other students, we were extremely badly hung over (more so than usual) and we had revised topics which did not come up on the paper. We know that if we took an exam for the same subject again we might expect to do better next time, more accurately reflecting our ability. This is because we would expect these sources of error (failures to record our true ability) to be less likely to *all* co-occur next time around. Similarly, if we were very lucky, the exam might only contain questions on the topics we had revised and we might be fortunate enough to sit the exam on the only day of the year when we were not hung over and everybody behaved themselves in the exam hall. This time we might get a mark that somewhat overstated our true ability in the subject. However, we probably would not expect to be so lucky if we took a similar exam again without further revision.

Across a sample of people, those with mid-range scores are likely to be about equally influenced by these errors (inflating and reducing scores) so they would cancel out on average, leading to no systematic bias in our experiment. People at the extremes, however, are *less likely* to score more extremely on being retested as some of those who had extreme scores at pre-test will have done, because their scores had already been inflated (or reduced) by chance factors or errors unrelated to ability. Since extremely large errors are relatively less likely than moderate size errors, two consecutive large errors in the same direction are very unlikely. This means that post-test scores will tend towards the population's mean score.

For quasi-experiments, this is a particular problem when the treatment group has been selected because of the participants' low scores on the dependent variable (e.g. selecting people with poor computing skills for the peer teaching method). The simplest way to guard against this (though easier said than done) is to ensure that your control group is also drawn from the pool of extreme scorers. The ethics of denying an intervention to children who are particularly bad at computing are clearly an issue here. The problem is also more likely to influence results if your dependent measure has low test–retest reliability. The less reliable the measure (i.e. the more error-prone it is) the more there is likely to be regression to the mean.

Contamination effects occur when, despite your best efforts, the treatment and control groups influence each other in some way. This can often be quite subtle and difficult to detect. In the computer skills study it might happen that a group of keen parents on hearing about the good peer teaching going on at another local school start an after-school club where their children engage in peer teaching around a computer. Thus, although their children are in the control group they may actually be experiencing the treatment, thus leading to potential ambiguity in interpreting the results if the children start to improve to the same levels as the treatment group. Contamination effects are a big problem for studies evaluating health treatments where participants may want to seek additional treatments on top of the one being studied. You might want to evaluate the efficacy of a new cognitive therapy intervention for depression and have a treatment group and a normal drug-treatment control group. Problems arise if those in your normal drug-treatment group also decide to seek a talking therapy from someone else or if your treatment group members even seek out other sources of anti-depressant medication.

Finally, history effects can affect the validity of NECG studies. If some event, in addition to the treatment intervention, occurs between pre-test and post-test in one group only, then it is not clear what any group differences at post-test should be attributed to. For example, an evaluation of a persuasive campaign to promote commuting to work by urban railways in different cities may be invalidated if the 'treatment' city suffers from road travel chaos caused by unanticipated roadworks on the main commuter routes during the period of the study. People may flock to the trains but only because driving to work (their preferred method) was nearly impossible on the test days. You should be aware that all these effects can work to enhance group differences *or* to obscure them.

4.5 TIME SERIES DESIGNS

Time series designs involve having only one sample but taking measurements of the dependent variable on three or more occasions. Such designs are sometimes referred to as interrupted time series designs as the treatment intervention 'interrupts' an otherwise seamless time series of observations. Figure 4.2 gives an illustration of some hypothetical time series data.

As you can see, the main feature that you are looking for when collecting time series data is that the only substantial change in scores coincides with the intervention. The virtue of such a design is that it is relatively less likely that short-term historical events (i.e. history effects) will either (a) co-occur with the treatment and/or (b) have a lasting effect over time. It is also unlikely that small differences pre- and post-intervention will be maintained if the treatment really has no effect. Any maturation effects should be reflected in gradual trends in time series data and not in radical changes occurring at the same time as the intervention.

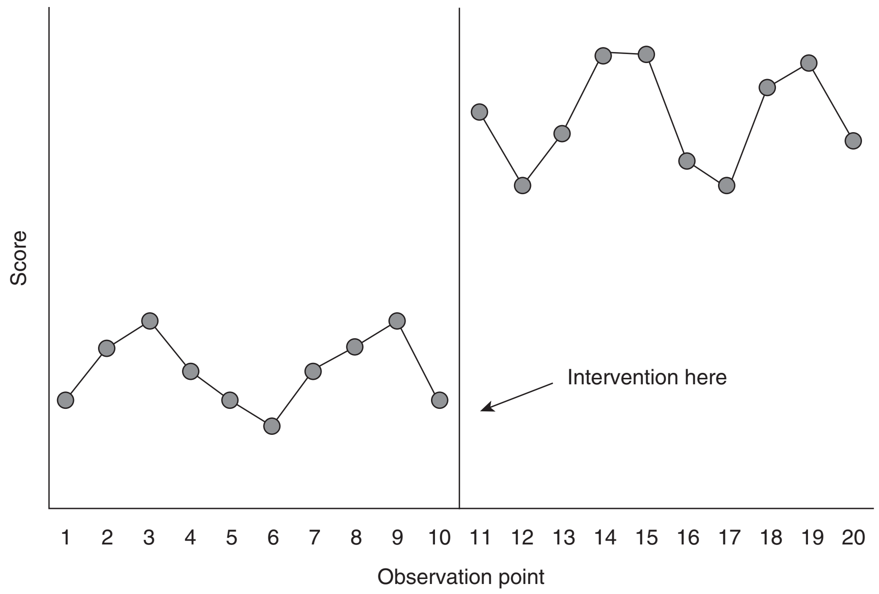

Figure 4.2 Example time series data

For time series studies to work well, multiple data collection/observation points are required. It is difficult to detect trends of any kind with just three observation points so, where possible, opt for as many observation points as is realistic but pay due regard to participant fatigue, boredom and irritation.

4.5.1 Problems with time series designs

Time series studies potentially suffer from the threat of testing effects to their validity. As these studies, by definition, require repeated administration of the same dependent measures, there is a tendency for people to gradually do better as time goes on. This is a separate phenomenon from maturation effects as testing effects arise out of familiarity with the measurement procedures. When presented with a novel test, for instance, we usually do not know what is required and may be anxious about our performance. Repeated exposure to the test and growing familiarity should reduce these anxieties and allow us to perform better. It is also possible that respondents might come to know what they are being asked about and develop more efficient answering strategies, allowing them to respond more quickly. This is especially a problem where measurements are timed.

The net impact of testing effects is that, if the *magnitude* of the treatment effect itself is small, it may get swamped by the testing effects. If the size of the treatment effect is relatively large there will be little problem in determining that the treatment actually had an effect.

Another potential problem concerns instrumentation effects. This refers to changes in the accuracy or reliability of measurements over time. One good example would be the reporting of crimes. Over time the likelihood of reporting (and the police recording) crimes changes as a function of changes in the social representation of the crimes rather than their frequency *per se*. What may have been regarded as common assault in the past may come to be seen as a racially motivated attack in more enlightened times. Similarly, women are now encouraged to report sexual attacks and the social opprobrium that used to follow a claim of rape is now somewhat reduced, though nonetheless still present. What this is really about is a change in the way the measures are taken and their relative accuracy. Studies that involve measures taken by observers are particularly at risk from instrumentation effects as observers learn how to use the coding schedule more efficiently or, more likely (and worse), become fatigued by the schedule and attempt their own reinterpretation of it.

Subject or participant mortality refers to the loss of participants from your study over time. Time series studies, especially those that cover long periods, are prone to participant mortality problems which are usually outside the experimenter's control. Some participants may indeed die during the study, but it is more normal that some will drop out through boredom or a lack of interest or perhaps because they move house. If you do not have a large sample to start with you run the risk that you will have too few people left at the end of the study to enable you to draw any reliable conclusions at all.

Participant mortality would not be such a great problem were it a truly random event. However, reasons for leaving that are related to the nature of the study (e.g. a lack of interest in the research topic or the intrusive nature of the measures) can lead to a situation where the surviving sample becomes progressively more biased in favour of showing that the treatment works. Say you were trying to evaluate the effect of a local waste recycling advertising campaign and had started regular assessments of how much waste people recycled. Even if you started with a fairly representative sample of the population, you might well find that by the time you had started the adverts and were collecting post-intervention observations, only environmentally committed people were still ready and willing to help you with the project. In all likelihood, your estimates of average post-intervention waste recycling behaviour would be considerably higher than the pre-intervention average, but this would be mainly due to sample mortality rather than the effect of the adverts.

Careful mapping of sample survivors' pre- and post-intervention behaviour would overcome this problem, but this is naturally a rather unsatisfactory solution since such a campaign was presumably intended to change the behaviour of the less environmentally committed people who were lost to the study. Needless to say, strenuous efforts have to be made to maintain the sample.

4.6 TIME SERIES WITH NON-EQUIVALENT CONTROL GROUP DESIGNS

Many of the problems associated with time series and NECG designs are neatly overcome by the combination of the two approaches in the time series with non-equivalent control group (TSNECG) design, sometimes also called the multiple time-series design. An extended series of data collection points are used with both the treatment group and the non-equivalent control. The key advantage of the TSNECG design is that you should be able to tell both whether a treatment has an effect compared with a control group and that the effect only occurs at a point after the introduction of the treatment. It helps to rule out many of the individual threats to validity outlined previously.

Figure 4.3 illustrates what we would hope to find if there really was a strong treatment effect. It shows there is variability in scores over time and there appears to be a gradual improvement in scores in the control group, potentially via testing, instrumentation or maturation effects; however, the post-intervention scores for the treatment group are considerably higher than for the controls suggesting that there really was an effect of the treatment intervention.

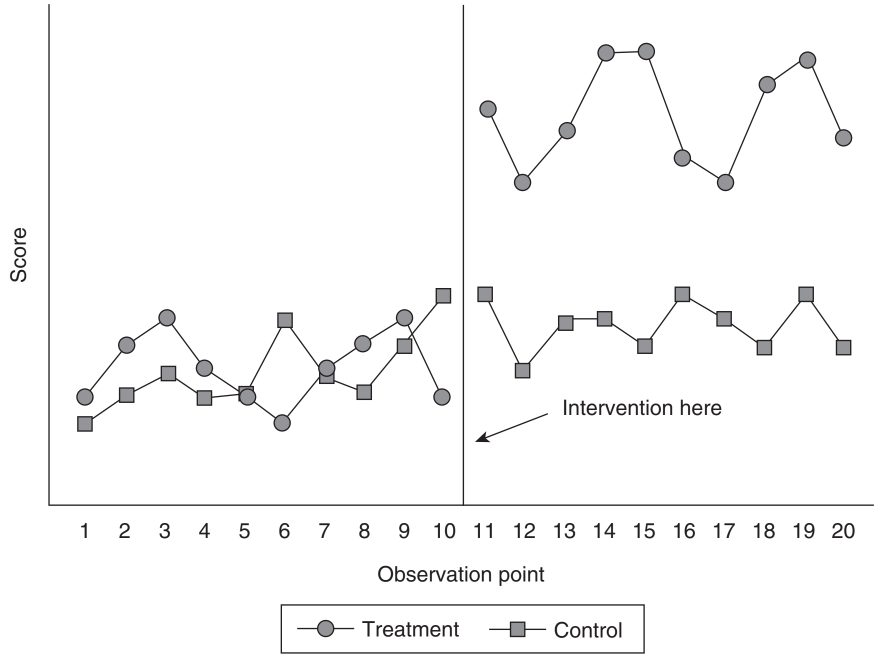

Figure 4.3 Time series with non-equivalent control group

4.6.1 Problems with TSNECG designs

The price to be paid for minimising so many threats to validity is all-round increased cost and the need to study many more people. This is not a problem when conducting research on existing archival data but may be a serious problem if you intend to collect fresh data.

Differential sample mortality in the two groups can be a problem. If people who are somewhat apathetic to the study are differentially more likely to be lost from one group than the other, then group differences may be artificially enhanced or constrained. It is also possible with studies that last for some time that the control group will become exposed to, or aware of, the treatment. People in the two groups may mix and discuss the intervention, and control group members may either seek the treatment for themselves or withdraw from the study through becoming aware that they may never be exposed to the treatment or intervention.

Sometimes, merely being aware of the existence of a 'problem' that needs treating may change behaviours of control group members. If control group members come to feel that they are being deliberately disadvantaged in some way they may choose to perform less well when measurements are taken. This may be a serious problem if researchers are heavy-handed and insensitive in the way they interact with people. Alternatively, control group members may compensate for not receiving the treatment by trying harder to perform well. This is called compensatory rivalry and would serve to obscure true treatment effects.

TSNECG designs are not immune to the other threats to validity discussed earlier, especially if the magnitude of the treatment effect is weak and the variability between scores on successive observations is relatively high. In common with the single-case designs (see Box 4.1), detecting a treatment effect is easiest when it is possible to establish a fairly clear-cut stable baseline in both the control group and the treatment group prior to the intervention. As with true experiments, it may be necessary to increase sample sizes substantially in order to provide the necessary statistical power to detect these weak effects.

4.7 MODIFICATIONS TO THE BASIC DESIGNS

The basic designs described here are really the tip of the iceberg in terms of possibilities. With NECG designs there is no necessity to have only two treatment conditions (treatment and control). It is possible to have many different levels of the treatment or combinations of treatments in one design. For example, we might extend the computer skills example to include a control (traditional teaching) group, a group that had two periods a week of peer teaching and one that had four per week. In fairness to traditional methods of teaching, we might also divide the control group into one that had two periods per week and one that had four periods of

Box 4.1 Single-case designs

Usually researchers are urged to seek out large samples to increase their confidence in the conclusions they draw from a study but it is perfectly possible to conduct meaningful experiments on single cases. The most common single-case design is the A-B-A design which shares many of the characteristics of the time-series design discussed in this chapter.

The A-B-A design is the best-known single-case experimental design in which the target behaviour or response is clearly specified and measurements are carried out continuously throughout three phases of the experiment: A, B and A again. The first occurrence of Phase A is the baseline phase during which the natural occurrence of the target behaviour or response is monitored; in Phase B the treatment/intervention is introduced. To increase our confidence that the treatment in Phase B is responsible for any changes we see, the treatment is then removed and responses monitored in what amounts to another baseline Phase A. A hypothetical example of an A-B-A design is illustrated in Figure 4.4.

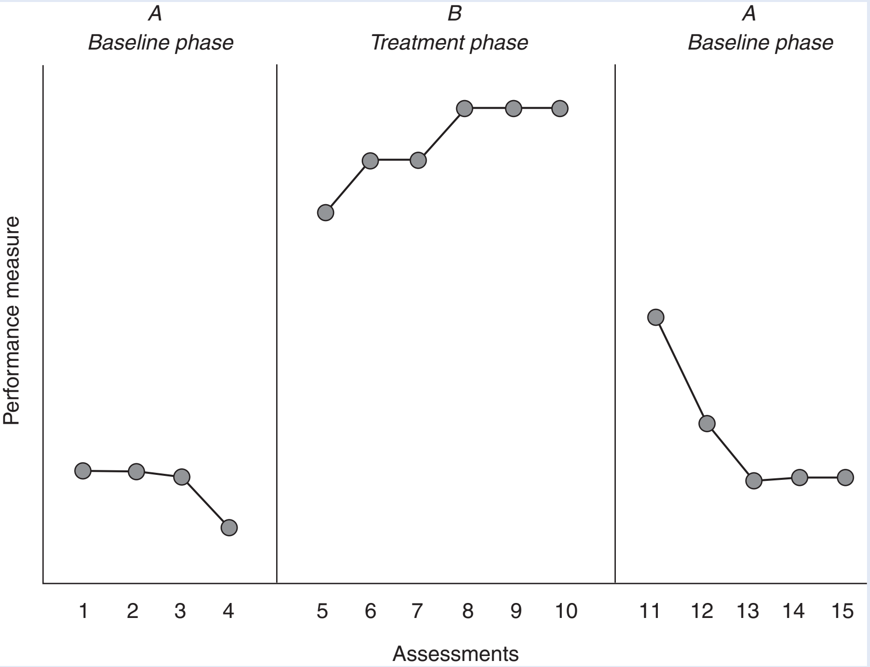

Figure 4.4 Example of an A-B-A design

There is however an important reservation concerning the clinical application of the A-B-A design; this is that it may not be possible to tell whether any behaviour change

Box 4.1 (Continued)

that occurs following onset of the treatment results from the treatment *per se* or from changes that are part of the recovery process that would have happened even without the treatment. This issue is particularly problematic when there is only weak evidence of an experimental effect; that is, only a slight improvement is seen. One way of overcoming this problem is to use a control variable. A control variable would be another aspect of behaviour which would be as susceptible to the effects of recovery as the experimental variable, but is not thought to be something that will be influenced by the treatment. If the effects found following treatment were due to naturally occurring, non-treatment-related recovery then the curves for the treatment and control variables should be parallel.

A-B-A designs are most commonly used in clinical settings where clinicians are interested in finding out whether a treatment intervention will work for a particular patient, usually with a relatively unique combination of problems (treatments for comparatively common conditions are usually tested using true experiments in the form of randomised control trials; see Chapter 3). The A-B-A design presents some fairly obvious ethical problems, as a potentially valuable treatment is systematically being removed from someone who might benefit from it. To deal with this, many variations on the A-B-A design have been suggested such as the A-B-A-B design where the study finishes with a treatment phase which can then be extended beyond the end of the study if the treatment works, but a phase of withdrawal still allows an opportunity for the efficacy of the treatment given in the B phases to be evaluated. There are other variations on the A-B-A-B design; for example, having multiple treatment and baseline phases (A-B-A-B-A-B-A-B-A-B) or incorporating another treatment (A-B-A-C-A-B-A-C).

traditional teaching. Clearly, this new design is much more useful to curriculum developers since it not only tells us whether peer teaching is better than traditional methods, but also whether spending more time on computing yields worthwhile increases in skill level. Assuming we had enough schools prepared to help, we could even add a group that gets both traditional and peer teaching for a total of four periods.

Sometimes concerns about testing effects may lead us to believe that post-test measures will be unduly influenced by people having completed the pre-test. An example might be of a knowledge test with a short period between pre- and post-test. In such a situation we might expect people to remember the items, thus inflating the apparent power of any intervention. It is also often the case that merely asking people about some aspect of their lives changes their behaviour in that domain. For instance, merely asking about your waste recycling activities might make you think that you ought to recycle more waste. Somebody showing interest in your behaviour may change it. This is called the Hawthorne effect after the electricity plant in Illinois where the phenomenon was first formally described in studies on attempts to enhance worker performance (Roethlisberger and Dickson, 1939). It is possible to get over both sorts of problem by using separate pre- and post-test samples so that different individuals take the pre- and post-tests. This approach is only sensible if you have a large pool of people from which to draw your samples and you can draw them by some fairly random procedure.

For time series designs and TSNECG designs it is possible to adopt treatment withdrawal designs. These involve intervening with the treatment and then, at a later point, withdrawing it and observing a subsequent fall in scores on the dependent measure. This approach works best when the treatment is not expected to have a lasting effect on the dependent variable and has to be 'maintained' in some sense for the effect to be shown. An example might be to evaluate the effectiveness of camera-based speed checks on stretches of road. Speeds could be monitored surreptitiously for some period before erecting the camera systems then, after a period with the cameras in place, they could be removed to see if speeds gradually increased in their absence. The cameras could be re-erected later to see if speeds fell again.

4.8 CONCLUSION

With all the potential problems associated with each quasi-experimental design, you might be thinking that they are too fraught with difficulties to make them worthwhile. The difficulties, however, are inevitable whenever you forego experimental control in order to do research outside the laboratory. What I hope to have shown you is that there are some rigorous methods available and, while they will not always lead you to unambiguous answers to your research questions, they do at least flag up the likely threats to validity. If you know where potential interpretative problems lie then you can address them and make some estimate of the likely impact these could have had on the results of your study. Quasi-experiments, providing they are conducted with due care, can be the most powerful available means by which to test important hypotheses.

4.9 EXERCISES

1 Design your own pre-experiment and think about what you would have to do to make it into a true experiment or a quasi-experiment.
2 Take a look at the week's newspapers and pick out the stories where something has been evaluated – there are usually lots of these in the 'quality' press. For each one, what kind of design was used? If you cannot tell, what information do you need in order to make the judgement? What are the likely threats to the validity of the conclusions that were reached?

4.10 DISCUSSION QUESTIONS

1 How far does the lack of control over allocation of participants to experimental conditions undermine the validity of findings of quasi-experiments?
2 Are pre-experiments essentially an unethical waste of participants' time?

4.11 FURTHER READING

The classic text in this area is Donald Campbell and Julian Stanley's (1966) *Experimental and quasi-experimental designs for research*. This is a very short book of only 70 pages which had first appeared as a chapter in Gage (1963) and it is the place where quasi-experimental designs were first comprehensively explained. William Shadish, Thomas Cook and Donald Campbell produced a more detailed text called *Experimental and quasi-experimental designs for generalized causal inference* in 2002. This contains discussions of the major designs and a few more, as well as information about the appropriate statistical models to be used with each design. For single-case studies John Todman and Pat Dugard's *Single-case and small-*n *experimental design* (2001) expands on the examples given here.