Evaluating and conducting intervention studies: a guide for speech and language therapists

Dorothy V.M. Bishop and Paul A. Thompson

2019-04-01

Contents

Preface

Why did I write this book?

Methods for evaluating effectiveness of interventions have been developed in the field of medicine, where practitioners who wish to run clinical trials typically have access to expert statisticians and methodologists. However, in many fields, professionals with expertise in administering interventions have little or no training in research design and statistics.

The idea for this book was prompted by attending a conference on interventions for children with developmental language disorders. As the meeting there were many committed practitioners who were passionate about trying to improve outcomes for these children, and who wanted to engage in studies to show that their interventions were effective. However, all too often they did studies in a way that could not show the effect they were interested in, and conclusions about the benefits of their interventions seemed over-optimistic. I realised that the dedication and skill of these practitioners was not matched by a deep understanding of the difficulties of intervention research; in particular, the clinical training of these professionals did not usually include adequate instruction in basics of research methodology and statistics. It seemed, therefore that there was a need for a basic text that would explain the pitfalls of intervention research, as well as providing a template for good practice in the evaluation and design of intervention studies.

As we shall see, demonstrating that an intervention has an impact is much harder than it appears at first sight. There are all kinds of issues that can arise to mislead us into thinking that we have an effective treatment when this is not really the case. Much of the attention of methodologists has focused on how to recognise and control for unwanted factors that can affect outcomes of interest. But, as a psychologist, I am also particularly interested in our own human biases that can be just as important in leading us astray. Good, objective intervention research is vital if we are to improve the lot of those we work with, but it is really difficult to do well, and to do so we have to overcome our natural impulses to interpret evidence in biased ways.

Who is this book for?

Although the inspiration for the book came from interactions with speech and language therapists, and the illustrative cases are from that discipline, the basic principles covered here are relevant for any field where a practitioner aims to influence outcomes of those they work with. This includes those working in professions allied to medicine and education.

What is covered?

My main goal is to instil in the reader awareness of the numerous sources of bias that can lead to mistaken conclusions when evaluating interventions. Real-life examples are provided with the aim of providing an intuitive understanding of these issues.

I expect that many readers will have little or no background in statistics. Lack of statistical training is a massive obstacle to practitioners who want to do intervention research: it not only makes design and analysis of a study daunting, but it also limits what a potential researcher can take from the existing literature. This book should be seen as complementing rather than substituting for a technical introduction to statistics. Many readers may be reluctant to study statistics in more depth, but it is hoped that the

account given here will give them confidence to approach statistics in the published literature with a more critical eye, to recognise when the advice of a professional statistician is needed, and to communicate more effectively with statisticians.

Having said that, I hope that at least a subset of readers will consider engaging with the exercises that I have included using the programming language R. These are optional and the book has been written to be comprehensible without needing to learn any programming skills. But I think you will take more away from it if you do learn some 'coding' – as the experts call it.

R has the advantage of being free to download, and there is a huge community of R users, which means that you can Google any question you might want to ask about it, and you are likely to find that someone has answered it. R is not, however, easy to learn, and it can be off-putting on first encounter. The huge advantage of R over more familiar and user-friendly statistics packages is that is that it is fairly easy to simulate data – that is, to generate artificial data with particular properties. For example, you could generate fictitious results from two groups - one who had an intervention and one who didn't. That probably seems like an odd thing to want to do – surely what you need is real data! The reason for doing it should become more apparent as you proceed through the exercises. With simulated data you can see the effects of the various biases that we will consider. For instance, you can see how easy it is to get a false positive 'significant' result if you run a large number of statistical tests on the same sample. And you can see how it is possible to mistakenly conclude that an intervention is ineffective if you use too small a sample size in a study.

Intervention research is a branch of science, and you can't do good science without adopting a critical perspective – to the research of yourself as well as others. I hope this book will make it easier to do that and so to improve intervention research in speech and language therapy as well as other fields.

Chapter 1

How can we know if we've made a difference?

Anthony was a 60-year-old builder who suffered a stroke that left him paralysed on his right side and globally aphasic (with difficulties producing and understanding language). He was discharged home after three weeks in hospital, by which time he recovered the ability to walk and talk, but still had severe word-finding problems. He received weekly sessions with a speech-and-language therapist (SLT) for 6 weeks, after which his word-finding difficulties had reduced markedly. He is full of praise for the SLT and says she made a huge difference.

At two years of age, Tina's speech was markedly delayed. She had an expressive vocabulary of just ten words (mummy, daddy, doggie, water, more, want, juice, milk, bread and bear), and her repertoire of speech sounds was limited, so what she said was often not intelligible to strangers. She was socially engaged and an assessment showed that she had age-appropriate understanding of what others said to her. The SLT worked with Tina's mother to encourage her to talk about the toys Tina was playing with and to repeat and expand on her utterances. The mother said she found this extremely useful and it transformed her interactions with her daughter. Six months later, there was a dramatic improvement in Tina's expressive language: her speech was much clearer and she was talking in 2-3 word utterances.

A teaching assistant works in a primary school in an area of high social deprivation. She has worked with the school's SLT to develop a language intervention programme with a class of 5-year-olds that involves regular group sessions of story-book reading with an emphasis on developing the children's vocabulary. A vocabulary test that was given at the start of the school term and again 3 months later shows that on average children know ten more vocabulary items after the intervention than they did at the outset. The class teacher was enthusiastic about the intervention and wants to roll it out to more classes.

These three vignettes illustrate the kinds of everyday problem confronting SLTs going about their daily work: language problems are identified, interventions implemented, and, in many cases, improvements are observed. But the thoughtful therapist will have a niggling doubt: yes, in each case we see an improvement in language skills, but would this have occurred anyway? People with aphasia often recover over time, late-talkers turn out to be 'late-bloomers' and children's vocabulary grows as they get older.

Readers might wonder whether we should worry. After all, in each case, the SLT used their professional judgement to intervene and an improvement in language was seen. So does it matter whether that improvement would have occurred anyway?

I take a very firm view on this question: it matters enormously, for four reasons.

- First, we owe it to those who receive our interventions to apply due diligence to ensure that what we are doing is evidence-based and unlikely to do harm – which in the best case may involve wasting people's time or money, and in the worst case could cause emotional damage. The purity of the motives of a practitioner is not sufficient to ensure this – they have to have the skills and willingness to consider the evidence dispassionately.
- If SLTs want to be taken seriously, then they need to show that what they do is evidence-based: otherwise they are no better than purveyors of dubious alternative health cures, such as aromatherapy or iridology.
- Third, someone – often the taxpayer – is paying for the SLT's time. If some interventions are ineffective, then the money could be better spent elsewhere.
- Fourth, if a profession relies purely on traditional practice to determine which interventions to use, then there is no pressure to develop new interventions that may be more effective.

Showing that an intervention works in an individual case is very hard – especially when dealing with a condition that fluctuates or is self-limiting. In later chapters we shall consider how we can make use of group studies to evaluate interventions, and how single case designs can sometimes give greater confidence that a real effect has been achieved. But first, it is important to recognise the range of factors that conspire to make it difficult to answer the question 'Did I make a difference?' To do this, we need to understand about random and systematic change.

1.1 The randomness of everything

Anyone who has tried to diet will be aware of how weight can fluctuate: a fact used to comic effect in the novel Bridget Jones' diary, where the protagonist's mood soared and plummeted depending on whether she was a pound heavier or lighter than the previous week. Some of the factors affecting what the scales say may be systematic and depend on calories ingested and worked off, but some will be fairly random: how you stand on the scales, whether you had a cup of tea prior to weighing, and whether the floor is level may affect the reading. And a different set of scales might give a different result. Figure 1 shows a notional weight chart for Bridget, who is obsessively measuring herself daily. This plot was generated from random numbers – something we will learn how to do in Chapter x. For now, however, the simple point I want to make is that all measures will show some fluctuation. This is variously referred to as noise, measurement error or random error in the intervention literature.

When we evaluate an intervention, we need to identify measures that can be used to assess the outcome. All too often, insufficient attention is given to the specific measures that we select, but these can have a dramatic influence on whether or not we find a treatment effect. Figure 2 shows two putative measures of a child's vocabulary, for a simulated situation where a vocabulary intervention does have a positive impact. In effect, what we are trying to do when we assess a treatment is to sift out the systematic effect of treatment from the background noise. It is obvious that this will be much harder if we have a measure that hops around a lot from one measurement occasion to the next – a 'noisy' measure.

One bad consequence of noisy measures is that they can show big changes from one occasion to another which, in individuals, might give a misleading impression of genuine treatment effects. Figure 3 shows simulated data from 10 people using the noisy measure when there is no true effect. We might be tempted to conclude that there is a subset of people who respond to the treatment – those shown in red in the figure. If we were to just ignore the remainder of the group, we could give a quite false impression of effectiveness of intervention. Unfortunately, this sometimes happens, when people fail to recognise that some gains in a measure could just be due to chance, and conclude instead that there is a subset of 'responders' and another subset of 'treatment resisters'. It is possible to design a study to avoid making this basic error, but you need to be aware of it to defend against it.

1.2 Systematic error

An even greater danger for the unwary researcher is a failure to take into account systematic error. This refers to the fact that there may be changes in the people we are intervening with that would occur whether or not they had the intervention. The three examples that I gave at the outset all illustrate this type of error: with many kinds of brain injury, including stroke, there is very rapid recovery immediately after the injury – thought to be related to basic physiological changes such as reduction of swelling – followed by a longer period, that can persist for months or years, during which recovery continues, albeit at a slower pace. But the amount and duration of such recovery can vary markedly depending on factors such as the age of the person and the nature, location and extent of the injury. The second example, of Tina, illustrates another case – late talkers – where substantial spontaneous improvement can occur. Despite a growing number of prospective studies of late talkers, we are still rather bad at predicting which children will be 'late bloomers', who turn out fine without any special help, and which will go on to have more persistent, long-term problems. Finally, we would expect children to gain new vocabulary as they get older: the difficulty is knowing just how much change it is reasonable to expect over a 3-month period as a consequence of age.

In all three cases, therefore, we are not only trying to detect a signal from among noise – i.e. an effect of intervention using measures that inevitably contain random error: we also want to know whether any signal we see is due to the intervention that was provided, or whether it is just part of natural change. This can be extremely hard to do, but methods have been developed that provide a rational approach to addressing the question.

It isn't always this difficult. Some conditions are more stable and pose less of a problem of spontaneous recovery. But, in my experience, the commonest error in the field of SLT intervention research is a failure to take into account systematic error, and so I will be focusing most on that. And methods that are designed to deal with systematic error are also valid for more stable situations, so if you can cope with those, then you will be in a strong position to evaluate intervention in other contexts.

Chapter 2

How to select an outcome measure

Suppose you want to evaluate the effectiveness of a parent-based intervention for improving communication in three-year-olds with poor language skills. You plan to assess their skills before the intervention, immediately after the intervention, and again six months later. The initial measurement period is known as the baseline – because it acts against as a reference point against which improvement can be measured.

There are many measures you could choose: the child's mean length of utterance (MLU), scores on a direct assessment such as preschool CELF, the parent's response on a language inventory such as xxx. You may wonder whether you should include as many measures as possible to ensure you cover all bases. However, as we shall see in chapter x, if you measure too many things, you run the risk of getting spurious results, so it is important to specify a primary outcome measure – the one you would put your money on as most likely to show the effect of interest, if you were a betting person.

The key questions you have to ask yourself are:

1. Is the measure reliable?
2. Is it sensitive?
3. Is it valid? i.e., does it measure what I want to measure?
4. Is it practical?

2.1 Reliability

You may be surprised to see reliability at the top of the list. Surely validity is more important? Well, yes and no. As shown in Figure x, there's not much point in having a measure that is reliable unless it is also valid. But a measure that is valid but not reliable is worse than useless in an intervention study, so I put reliability at the top of the list. (targets figure here)

So what is reliability? This has to do with the issue of random error or 'noise', which can be estimated by seeing how far scores on one testing occasion line up with scores on another occasion close in time (i.e. before we expect any change due to maturation or intervention). Quite simply you want a measure that reflects the child's underlying skill, where there is minimal influence from random, unwanted sources of variation. So if we have two occasions of measurement, they should be closely similar. The similarity can be quantified by a correlation coefficient, demonstrated in Figure x.

Let's illustrate this with mean length of utterance (MLU). Roger Brown's classic work first showed that in very young children this is a pretty good indicator of a child's language level (ref and ?plot). As children grow older, though, it become less useful, and it may be strongly influenced by the context in which language is sampled. Length of the language sample will also be important: if you just based a MLU estimate on ten child utterances, then the estimate will be less stable than if you took 50 utterances.

We can't say that MLU is inherently good or bad on the reliability front: it will depend on numerous factors. It may work well if assessed from a reasonable length of language sample collected in closely similar contexts. Reliability may be better for a clinical sample, where language skills are less variable, than for typically-developing children. The important thing is to find out as much

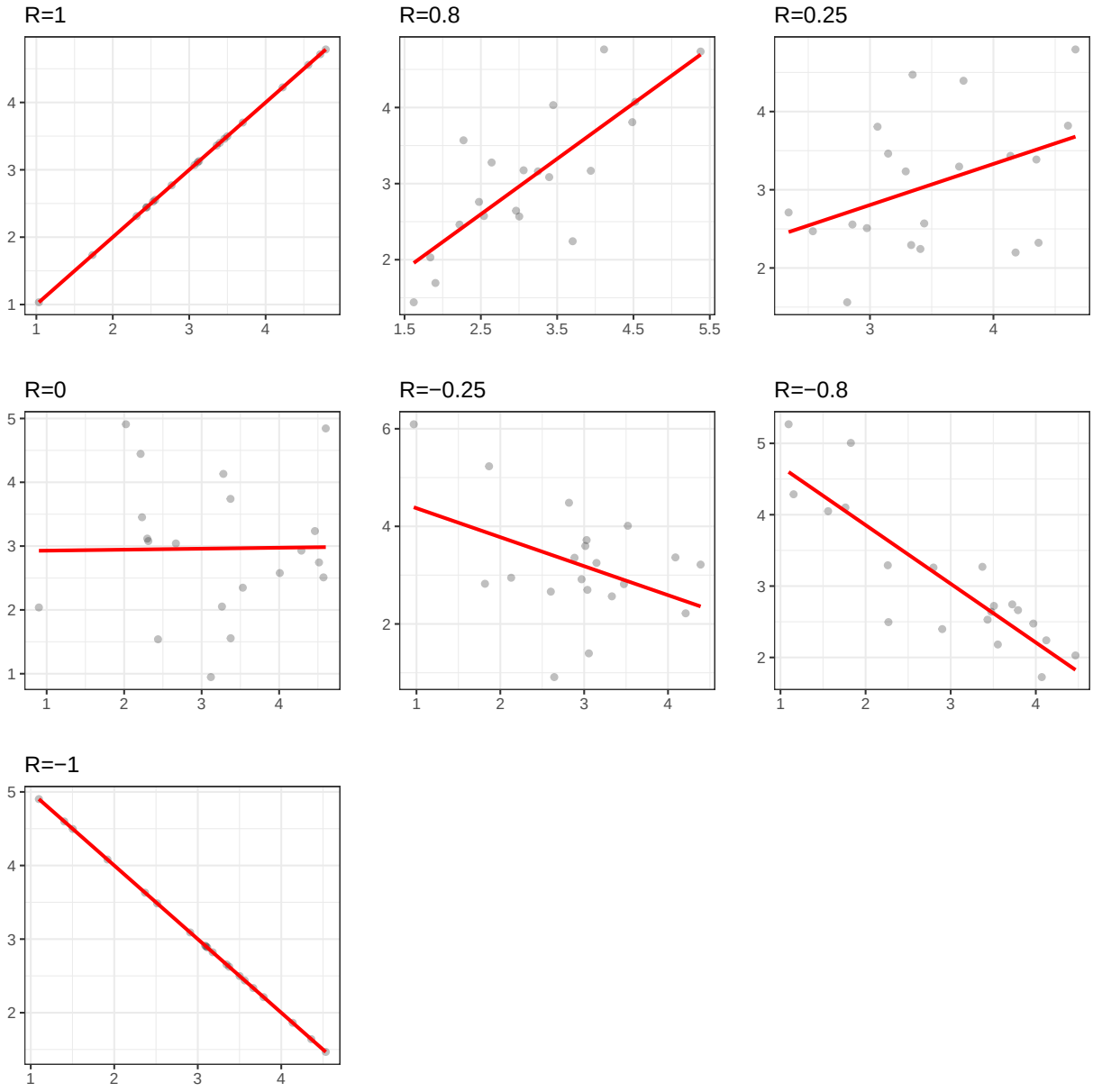


Figure 2.1: The correlation coefficient is a statistic that takes the value of zero when there is no relationship between two variables, one whether there is a perfect relationship, and minus one when there is an inverse relationship. If you draw a straight line through the points on the graph, then if all points fall exactly on the line, the correlation is 1, indicating that you can predict perfectly a person's score on Y if you know their score on X.

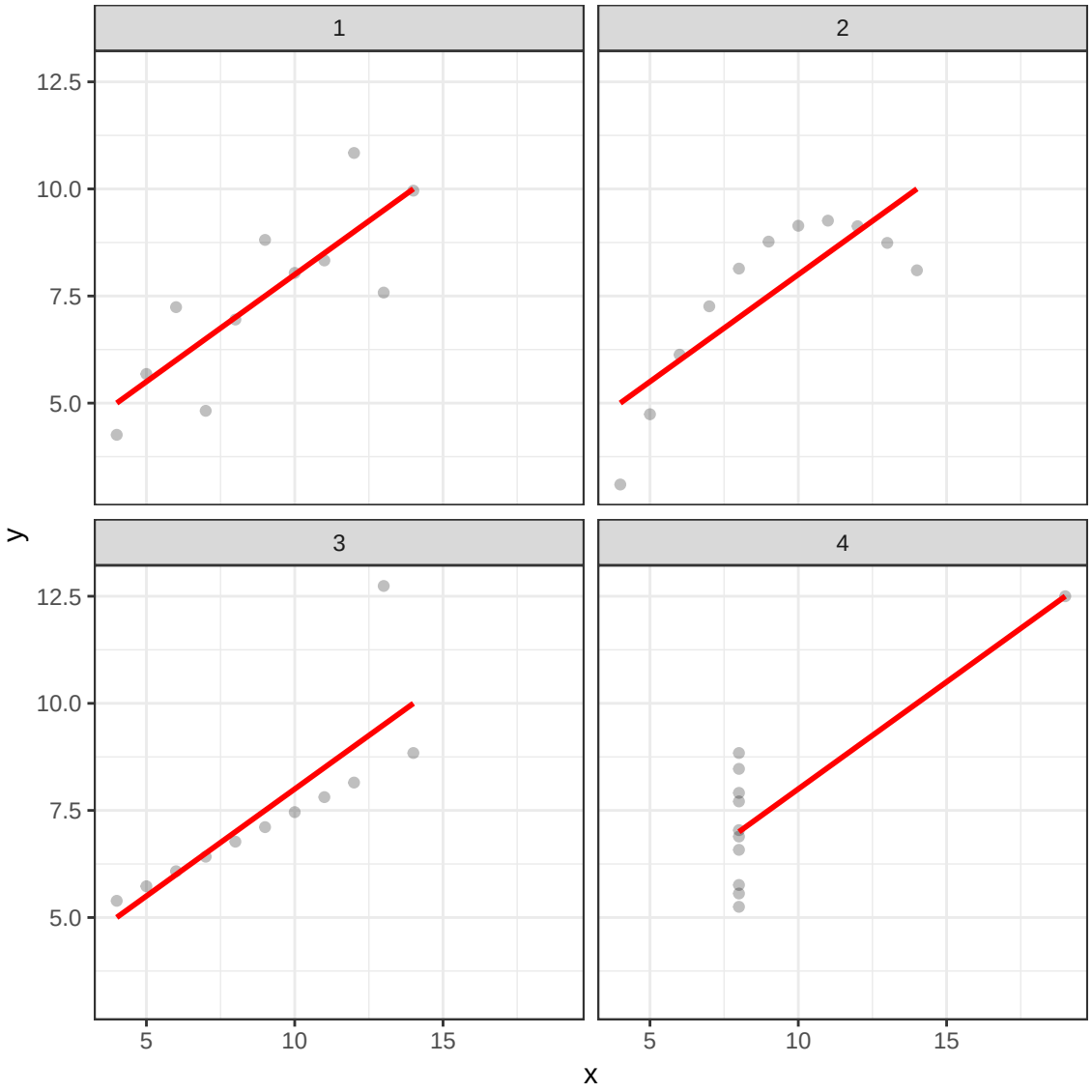


Figure 2.2: Anscombe's Quartet

Table 2.1: Different data same correlation

plot	cor(x, y)
1	0.8164205
2	0.8162365
3	0.8162867
4	0.8165214

as you can about reliability before committing to an outcome measure – or even better, do your own test-retest study with the measures of interest before embarking on a study.

So how reliable should a measure be? Most psychometric tests report reliability estimates, and a good test is expected to have test-retest reliability of at least .8. But be careful in interpreting such estimates, as you need also to consider the age range on which the estimate is based. Figure x shows how a test that has very poor reliability within 3-year-olds looks highly reliable when considered across the full age range from 3 to 8 years. This is because the index of reliability, the correlation coefficient, is affected by the range of scores. If your study is focused just on 3-year-olds, you really want to know how reliable it is just within that age range.

The topic of reliability is covered more formally in test theory (ref). This involves a mathematical approach that treats an observed test score as the sum of a 'true' effect (i.e. what you want to measure) plus random error. The lower the reliability, the greater the random error, and the harder it is to detect the true effect of intervention against the background of noise. Figure x illustrates this. (illustrative figure)

I have focused on test-retest reliability as this is the most relevant in intervention studies. If you plan to use the same measure at baseline and after intervention, then what you need to know is how much variation in that measure is likely to occur just by chance. There are other reliability indices that are sometimes reported with psychometric tests. In particular split-half reliability and internal consistency (Cronbach's alpha), both of which consider the extent to which a score varies depending on the specific items used to calculate it. For instance, we could assess split half reliability for MLU by computing it separately for all the odd-numbered utterances and the even-numbered utterances. Although this gives useful information, it is usually overestimates test-retest reliability, because it does not take into account fluctuations in measurement that relate to changes in the context or the child's state.

It is much easier to compute measures of internal consistency than to do the extra testing that is needed to estimate test-retest reliability, and many published psychometric tests only provide that information. Table x shows reliability estimates from some commonly used tests. (table

2.2 Sensitivity

Those who develop psychometric tests often focus on reliability and validity but neglect sensitivity. Yet sensitivity is a vital requirement for an outcome measure in an intervention study. This refers to the grain of the measurement: whether it can detect small changes in outcome. Consider Bridget Jones on a holiday to a remote place where there are no scales, just a primitive balance measure that allows her to compare herself against weights of different sizes (cartoon here). She would be unable to detect the daily fluctuations in pounds, and only be able to map her weight change in half-stone units. She could genuinely lose weight but be unaware of the fact.

Many standardized tests fall down on sensitivity, especially in relation to children scoring at the lower end of the ability range. It is customary to convert raw scores into scaled scores on these tests. This allows us to have a single number that can be interpreted in terms of how well the child is performing relative to others of the same age. But these often reduce a wide range of raw scores to a much smaller set of scaled scores, as illustrated in Table x. This means that a child could make substantial gains in raw score after intervention, but still come out with the same scaled score. (Cathy Adams; Elspeth McCartney; EEF)

For this reason, it is often recommended that raw scores be used for evaluating intervention effects. This is fine if the study involves a narrow age range, but it is more problematic when a wider range is used, because age will be a major determinant of test score. Indeed, that is one reason why scaled scores are often preferred: they allow us to compare children of different ages

on the same metric – i.e. a score that reflects how the child's performance relates – statistically – to that of a normative group of the same age.

In the case where a wide age range is used, raw scores can work well provided an appropriate statistical analysis is performed to take into account the age differences. We will cover this topic in chapter x.

Problems with sensitivity can also be an issue with measures based on rating scales. For instance, if we just categorise children on a 5-point scale as 'well below average', 'below average', 'average', 'above average' or 'well above average', we are stacking the odds against showing an intervention effect – especially if our focus is on children who are in the bottom two categories to start with. Yet we also know that human raters are fallible and may not be able to make finer-grained distinctions. Some instruments may nevertheless be useful if they combine information from a set of ratings.

Although we need sensitive measures, we should not assume that a very fine-grained measure is always better than a coarser one. For instance, we may be measuring naming latency in aphasic patients as an index of improvement in word-finding. It's unlikely that we need millisecond precision in the timing, because the changes of interest are likely to be in the order of tenths of a second at most. While there's probably no harm in recording responses to the nearest millisecond, this is not likely to provide useful information.

2.3 Measuring the right thing: validity

A modification of a popular adage is 'If a thing is not worth doing, it's not worth doing well.' This applies to selection of outcome measures: you could have a highly reliable and sensitive measure, but if it is not measuring the right thing, then there's no point in using it.

Deciding what is the 'right thing' is an important part of designing any invention study, and it can be harder than it appears at first sight. The answer might be very different for different kinds of intervention. I'll start with an issue that is particularly relevant to the first and third vignettes from chapter 1, word-finding intervention for aphasia, and the classroom-based vocabulary intervention

Generalisability of results: the concepts of far and near transfer The vignettes on word-finding intervention and vocabulary training illustrate interventions that have a specific focus. This means we can potentially tie our outcome measures very closely to the intervention: we would want to measure speed of word-finding in the first case, and vocabulary size in the second.

There is a risk, though, that this approach would lead to trivial findings. If we did a word-finding training with an aphasic patient using ten common nouns and then showed that they his naming had speeded up on those same ten words, this might give us some confidence that the training approach worked (though we would need appropriate controls, as discussed in later chapters). However, ideally, we would want to the intervention to produce effects that generalised and improved his naming across the board. Similarly, showing that a teaching assistant can train children to learn ten new animal names is not negligible, but it doesn't tell us whether this approach has any broader benefits.

These issues can be important in situations such as phonological interventions, where there may be a focus on training the child to produce specific contrasts between speech sounds. If we show that they master those contrasts but not others, this may give us confidence that it was the training that had the effect, rather than spontaneous maturation (see Chapter x), but at the same time we might hope that training one contrast would have an impact on the child's phonological system and lead to improved production of other contrasts that were not directly trained.

These examples illustrate the importance of testing the impact of intervention not only on particular training targets, but also on other related items that were not trained. As noted above, this is something of a two-edged sword. We may hope that treatment effects will generalise, but if they do, it can be difficult to be certain that it was our intervention that brought about the change. The important thing when planning an intervention is to think about these issues and consider whether the mechanism targeted by the treatment is expected to produce generalised effects, and if so to test for those. This is discussed further in chapter 4.

The notion of generalisation assumes particular importance when the intervention does not directly target skills that are of direct relevance to everyday life. An example is CogMed training, which is a computer-based intervention that has been promoted as a way of improving children's working memory and intelligence. The child plays games that involve visual tasks that tax working memory, with difficulty increasing as performance improves. Early reports maintained that training on these tasks led to improvement on nonverbal intelligence, as assessed by Raven's Matrices. However, more recent literature has challenged this claim, arguing that what is seen is 'near transfer' – i.e. improvement in the types of memory task that are trained – without any

evidence of 'far transfer' – i.e. improvement in other cognitive tasks. This is still a matter of hot debate, but it seems that many forms of 'computerised brain training' that are available commercially give disappointing results. If repeatedly doing computerised memory exercises only improves the ability to do those exercises, with no 'knock on' effects on everyday functioning, then the value of the intervention is questionable. It would seem preferable to use the time on training skills that would be useful in the classroom.

2.4 Functional outcomes vs test scores

In the second vignette we have an intervention where issues of far and near transfer are not relevant, as the intervention does not target specific aspects of language, but rather aims to modify the parental communicative style in order to provide a general boost to the child's language learning and functional communication. This suggests we need a rather general measure, and we are likely to consider using a standardized language test because this has the advantage of providing a reasonably objective and reliable approach to measurement. But does it measure the things that clients care about? Would we regard our intervention as a failure if the child made little progress on the standardized test, but was much more communicative and responsive to others? Or even if the intervention led to a more harmonious relationship between parent and child, but did not affect the child's language skills?

We might decide that these are the most important key outcomes, but then we have to establish how to measure them. In thinking about measures, it is important to be realistic about what one is hoping to achieve. If, for instance, the therapist is working with a client who has a chronic long-term problem, then the goal may be to help them use the communication skills they have to maximum effect, rather than to learn new language. The outcome measure in this case should be tailored to assess this functional outcome, rather than a gain on a measure of a specific language skill.

2.5 Subjectivity as a source of bias

In chapters x and x I will discuss various sources of bias that can affect studies, but one that crops up at the measurement stage is the impact of so-called 'demand characteristics' on subjective ratings. Consider, for a moment, how you respond when a waiter comes round to ask whether everything was okay with your meal. There are probably cultural differences in this, but the classic British response is to smile and say it is fine even if it was disappointing. We tend to adopt a kind of 'grade inflation' to many aspects of life when asked to rate them, especially if we know the person whose work we are rating.

In the context of intervention, people usually want to believe that interventions are effective and they don't want to appear critical of those administering the intervention, and so ratings of language are likely to improve from baseline to post-test, even if no real change has occurred. This phenomenon has been investigated particularly in situations where people are evaluating treatments that have cost them time and money (cognitive dissonance) but it is likely to apply even in experimental settings when interventions are being evaluated at no financial cost to those participating.

An example of this in the published literature comes from x et al who did a small-scale study to evaluate a computerised language intervention, FastForword (FFW). I will discuss larger evaluations of FFW in chapter x, but this study is noteworthy because as well as measuring children's language pre and post intervention, it included parent ratings of children's outcomes. There was a striking dissociation between the positive parent reports and the lack of improvement on language tests. Another example comes from a well-conducted trial of 'Sunshine therapy' for children with a range of neurodevelopmental disorders; here again we see that parents were very positive about the intervention, while objective measures showed children had made not significant progress relative to a control group.

Such results are inherently ambiguous. It could be that parents are picking up on positive aspects of intervention that are not captured by the language tests. For instance, in the Sunshine therapy study, parents reported that their children had gained in confidence – something that was not assessed by other means. However, there it is hard to know whether these evaluations are valid, as they are likely to be contaminated by demand characteristics.

Ideally we want measures that are valid indicators of things that are important for functional communication, yet are reasonably objective – and they need also to be reliable and sensitive! I don't have simple answers as to how this can be achieved, but it is important for researchers to discuss these issues when designing studies to ensure they achieve optimal measures.

2.6 Is it practical?

Intervention research is usually costly because of the time that is needed to recruit participants, run the intervention and do the assessments. There will always be pressures, therefore, to use assessments that are efficient, and provide key information in a relatively short space of time.

In my career I have occasionally been asked to advise people who are designing an intervention study, and I find that practicality is often very low on the list of topics that is discussed. A common experience is that the researchers want to measure everything they can think of in as much detail as possible. This is understandable: one does not want to pick the wrong measure and so miss an important impact of the intervention. But, as noted above, and discussed more in chapter x, there is a danger that too many measures will just lead to spurious findings. And each new measure will incur a time cost, which will ultimately translate to a financial cost, as well as potentially involving participants in additional assessment. There is, then, an ethical dimension to selection of measures: we need to optimise our selection of outcome measures to fulfil all the criteria of reliability, sensitivity and validity, but also to be as detailed and complex as we need, but no more.

My interest in efficiency of measurement may be illustrated with a vignette. When I started out in research, I was not involved in an intervention study, but I was embarking on a longitudinal study of 4-year-olds with developmental language disorders (Bishop & Edmundson, 1987). I took advice on what measures to use, and a common piece of advice was that I should take a language sample, and then analyse it using LARSP (Crystal et al.,xxxx), which at the time was a popular approach to grammatical analysis. Fortunately, I also had the chance to meet Catherine Renfrew, a retired SLT who had developed some language assessments for this age range – a sentence elicitation task, the Action Picture Test – and a narrative task – the Bus Story. As a practising therapist, Renfrew saw the need for short assessments that were easy to administer and interpret, and the tests she had developed fitted the bill.

I tried language sampling in my study but it seemed to provide little useful information in relation to the time it took to gather and transcribe the sample. Many of the children in my study rather little and did not attempt complex constructions. I found I could get more information in five minutes with the two Renfrew tests thank I could in 30 minutes of language sampling. Furthermore, I had more confidence in the test results, as all children were given the same task. When I came to do a grammatical analysis, I found that, after many hours of training in LARSP, analysing the results, and attempting to extract a quantitative measure from this process, I ended up with something that had a correlation of greater than .9 with mean length of utterance (MLU). The lesson I learned was that the measure needs to fit the purpose of what you are doing. I wanted an index of grammatical development that could be used to predict children's future progress. The Renfrew tasks provided to be among the most effective measures for doing that. A therapist working with a child might well find LARSP and language sampling useful for identifying therapy targets and getting a full picture of the child's abilities, but for my purposes this was far more detail than I needed.

There are other cases where researchers do very complex analysis in the hope that it might give a more sensitive indicator of language, only to find that it is highly intercorrelated with a much simpler index: use of xx index in nonword repetition is one example – this score, which takes into account the specific errors made in nonword repetition, is highly correlated with a score based on number of nonwords correct. In the domain of expressive phonology, it was only after I tried to develop an index based on analysis of phonological processes that I found that this was entirely predictable from a much simpler measure of percentage consonants correct.

A related point is that researchers are often tempted by the allure of the new, especially when this is associated with fancy technology, such as methods of brain scanning or eye-tracking. Be warned: these approaches yield masses of data that are extremely complex to analyse, and they typically are not well-validated in terms of reliability, sensitivity or validity! Even when high-tech apparatus is not involved, the newer the measure, the less likely it is to be psychometrically established – some measures of executive functioning fall in this category, as well as most measures that are derived from experimental paradigms. Clearly, there is an important place for research that uses these new techniques to investigate the nature of language disorders, but that place is not as outcome measures in intervention studies.

So, on the basis of my experience, I would advise that if you are tempted to use a complex, time-consuming measure, it is worthwhile first doing a study to see how far it is predictable from a more basic measure targeting the same process. It may save a lot of researcher time and we owe it to our research participants to do this due diligence to avoid subjecting them to unnecessarily protracted assessments.

2.7 Class exercise

Pick one of the three vignettes from Chapter 1 – or if you prefer, pick another intervention scenario – and make a list of possible outcome measures. For each one, evaluate its reliability, sensitivity, validity and practicality. At the end of this exercise, try to come up with a primary outcome measure that will be best for demonstrating an effect of intervention.

Chapter 3

Limitations of the pre-post design: biases related to systematic change

At first glance, assessing an intervention seems easy. Having used the information in chapter 2 to select appropriate measures, you administer these to a group of people before starting the intervention, and again after it is completed, and then look to see if there has been meaningful change. This is what I call a pre-post design, and it almost always gives a misleadingly rosy impression of how effective the intervention is. The limitations of such studies have been well-understood in medicine for decades, yet in other fields they persist, perhaps because they typically give results that look good.

Figure x shows some real data from a study I conducted, in which children were trained in a computerised game that involved listening to sentences and moving objects on a computer screen to match what they heard. The training items assessed understanding of word order, in sentences such as 'the book is above the cup', or 'the girl is being kicked by the boy'. There were two treatment conditions, but the difference between them is not important for the point I want to make, which is that we saw substantial improvement on a language comprehension test administered at baseline and again after the intervention in both groups. If this was the only data available, then we might have been tempted to conclude that the intervention was effective. However, we also had data from a third group who just had 'teaching as usual' in their classroom. They showed just as much improvement as the groups doing the intervention, indicating that whatever was causing the improvement, it wasn't the intervention.

To understand what was going on in this study, we need to recognise that there are several reasons why scores may improve after an intervention that are nothing to do with the intervention. These form the systematic biases that I mentioned in Chapter 1, and they can be divided into three kinds:

- Spontaneous improvement
- Practice effects
- Regression to the mean

3.1 Spontaneous improvement

We encountered spontaneous improvement in Chapter 1, where it provided a plausible reason for improved test scores in all three vignettes. People with acquired aphasia may continue to show recovery for months if not years after the brain injury, regardless of any intervention. Some toddlers who are 'late talkers' suddenly take off after a late start and catch up with their peers. And children in general get better at doing things as they get older. This is true not just for the more striking cases of late bloomers, but also for more serious and persistent problems. Children with autism spectrum disorder or severe comprehension problems do improve over time – it's just that the improvement may start from a very low baseline, so they fail to 'catch up' with their peer group.

Most of the populations that SLTs work with will show some spontaneous improvement which must be taken into account when evaluating intervention. Failure to recognise this is one of the factors that keeps snake-oil merchants in business: there are numerous unevidenced treatments on offer for conditions like autism and dyslexia. Desperate parents, upset to see their children

struggling, subscribe to these. Over time, the child's difficulties start to lessen, and this is attributed to the intervention, creating more 'satisfied cases' that can then be used to advertise the intervention to others.

As shown in Figure 3.1, one corrective to this way of thinking is to include a control group who either get no intervention or 'treatment as usual'. If these cases do as well as the intervention group, we start to see that the intervention is not as it seems.

3.2 Practice effects

The results that featured in Figure 3.1 are hard to explain just in terms of spontaneous change. The children in this study had severe and persistent language problems for which they were receiving special education, and the time lag between initial and final testing was relatively short. So maturational change seemed an unlikely explanation for the improvement. However, practice effects were much more plausible.

A practice effect, as its name implies, is when you get better at doing a task simply because of prior exposure to the task. It doesn't mean you have to explicitly practice doing the task – rather it means that you can learn something useful about the task from previous exposure. One of my favourite illustrations of this is a balancing device that came as part of an exercise programme called Wii-fit. This connected with a controller box and the TV, so you could see exercises and try specific tests that were shown on the screen. Users were encouraged to do the exercises to estimate their 'brain age'. When I first tried the exercises, my brain age was estimated around 70 – at the time about 20 years more than my chronological age. But just a few days later, when I tried again, my brain age had improved enormously to be some 5 years younger than my chronological age. How could this be? I had barely begun to use the Wii-fit and so the change could not be attributed to the exercises. Rather, it was that familiarity with the evaluation tasks meant that I understood what I was supposed to do and could respond faster and apply strategies to optimise my performance.

It is often assumed that practice effects don't apply to psychometric tests, especially those with high test-retest reliability. However, that doesn't follow. High reliability just tells you whether the rank ordering of a group of people will be similar from one test occasion to another – it does not say anything about whether the average performance will improve. In fact, we now know that some of our most reliable IQ tests show substantial practice effects persisting over many years. (Rabbitt studies). One way of attempting to avoid practice effects is to use 'parallel forms' of a test – that is different items with the same format. Yet that does not necessarily prevent practice effects, if these depend primarily on familiarisation with task format and development of strategies.

There is a simple way to avoid confusing practice effects with genuine intervention effects, and it's the same solution as for spontaneous improvement – include a control group who don't receive the intervention. They should show exactly the same practice effects as the treated group, and so provide a comparison against which intervention-related change can be assessed.

3.3 Regression to the mean

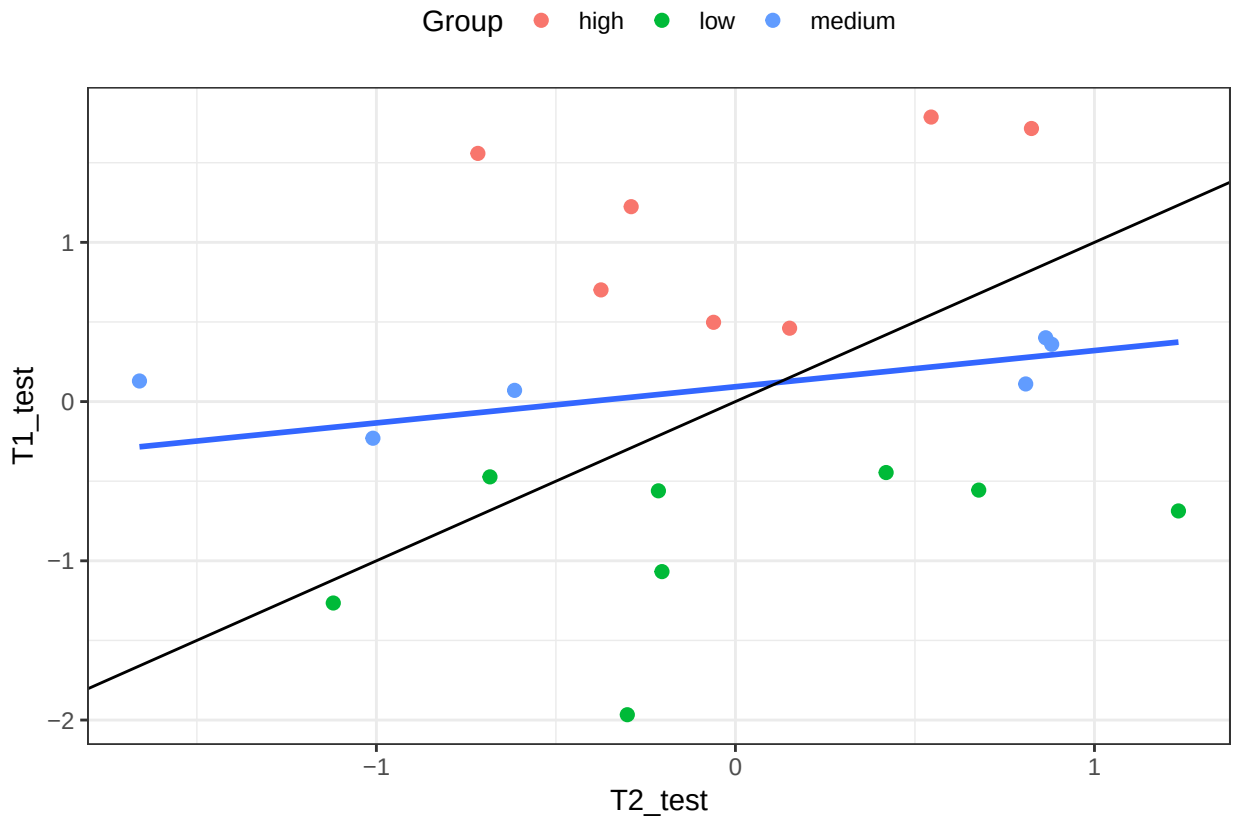
Spontaneous improvement and practice effects are relatively easy to grasp intuitively using familiar examples. Regression to the mean is quite different – it actually appears counter-intuitive to most people and it is frequently misunderstood.

It refers to the phenomenon whereby, if you select a group on the basis of poor scores on a measure, X, then the worse the performance is at the start, the greater the improvement you can expect on re-testing. The key to understanding regression to the mean is to recognise two conditions that are responsible for it: a) the same measure is used to select cases for the study and to assess their progress and b) the measure has imperfect test-retest reliability.

Perhaps the easiest way to get a grasp of what it entails is to suppose we had a group of 10 people and asked them each to throw a dice 10 times and total the score. Let's then divide them into the 5 people who got the lowest scores and the 5 who got the highest scores and repeat the experiment. What do we expect? Well, assuming we don't believe that anything other than chance affects scores (no supernatural forces or 'winning streaks'), we'd expect the average score of the low-scorers to improve, and the average score of the high scorers to decline. This is because the probability for any one person is that they will get an average score on any one set of dice throws. So that's the simplest case, when we have a score that is determined **only** by chance: i.e. the test-retest reliability of the dice score is zero.

Cognitive test scores are interesting here because they are typically thought of comprising two parts: a 'true' score, which reflects how good you really are at whatever is being measured, and an 'error' score, which reflects random influences. Suppose, for instance, that you test a child's reading ability. In practice, the child is a very good reader, in the top 10% for her age, but the exact number of words she gets right will depend on all sorts of things: the particular words selected for the test (she may know 'catacomb' but not 'pharynx'), the mood she is in on the day, whether she is lucky or unlucky at guessing at words she is uncertain about. All these factors would be implicated in the 'error' score, which is treated just like an additional chance factor or throw of the dice affecting a test score. A good test is mostly determined by the 'true' score, with only a very small contribution of an 'error' score, and we can identify it by the fact that children will be ranked by the test in a very similar way from one test occasion to the next, i.e. there will be good test-retest reliability. In other words, the correlation from time 1 to time 2 will be high.

groups	Freq
high	7
low	8
medium	6



Group	T1_Mean	T2_Mean
Low	-0.8776406	-0.0240830
Medium	0.1401478	-0.1218460
High	1.1349844	0.0110685

Simulated test scores for 21 children on tests varying in reliability. These figures show simulated scores for a group of 21 children on tests that vary in test-retest reliability. In each case individuals are colour-coded depending on whether they fall in the bottom (blue), middle (purple) or top (red) third of scores at time 1. The simulations assume no systematic differences between time 1 and 2 - i.e. no intervention effect, maturation or practice. Scores are simulated as random numbers at time 1, with time 2 scores then set to give a specific correlation between time 1 and 2, with no change in average score for the group as a whole.

Now suppose we select children because they have a particularly low score at time 1. Insofar as chance contributes to their scores, then at time 2, we would expect the average score of such a group to improve, because chance pushes the group average towards

the overall mean score. The left-hand panel shows the situation when reliability (i.e., correlation between time 1 and time 2 scores) is zero, so scores are determined just by chance, like throwing a dice. The mean scores for the blue, purple and red cases at time 1 are, by definition different (they are selected to be low, medium and high scorers). But at time 2 they are all equivalent. The upshot is that the mean at time 2 for those with initial low scores (blue) goes up, whereas for those that start out with high scores, the time 2 mean comes down.

The middle and right-hand panels show a more realistic situation, where the test score is mixture of true score and error. With very high reliability (right-hand panel) the effect of regression to the mean is small, but with medium reliability (middle panel) it is detectable by eye even for this very small sample.

The key point here is that if we select individuals on the basis of low scores on a test (as is often done, e.g. when identifying children with poor reading scores for a study of a dyslexia treatment), then, unless we have a highly reliable test with a negligible error term, the expectation is that the group's average score will improve on a second test session, for purely statistical reasons. In general, psychometric tests are designed to have reasonable reliability, but this varies from test to test and is seldom higher than .75-.8.

So regression to the mean is a real issue in longitudinal studies. It is yet another reason why scores will change over time. Zhang and Tomblin (2003) noted that we can overcome this problem by using different tests to select cases for an intervention study and to measure their improvement. Or we can allow for regression to the mean if our study includes a control group, who will be subject to the same regression to the mean as the intervention group.

3.4 Class exercise

Identify an intervention study on a topic of interest to you – you could do this by scanning through a journal, or by typing relevant keywords into a database such as Google Scholar, Web of Science or Scopus. If you are not sure how to do this, your librarian should be able to advise. It is important that the published article is available so you can read the detailed account. If an article is behind a paywall, you can usually obtain a copy by emailing the corresponding author.

Your task is to evaluate the article in terms of how well it addresses the systematic biases covered in this chapter. Are the results likely to be affected by spontaneous improvement, practice effects, or regression to the mean? Does the study design control for these? Note that for this exercise you are not required to evaluate the statistical analysis: the answers to these questions depend just on how the study was designed.

Chapter 4

Improvement due to nonspecific effects of intervention

4.1 Placebo effects, the Hawthorne effect and the Rosenthal effect

Most of us are familiar with the placebo effect in medicine: the finding that patients can show improvement in their condition even if given an inert sugar pill. There is much debate about the nature of placebo effects – whether they are mostly due to the systematic changes discussed in chapter 3, or something else. They are thought to operate in cognitive and behavioural interventions as well: communication may improve in a person with aphasia because they have the attention of a sympathetic professional, rather than because of anything that professional does. And children may grow in confidence because they are made to feel valued and special by the therapist.

This resembles the well-known Hawthorne effect, which refers to a tendency for people to change their behaviour as a result of being observed. The name comes from the Hawthorne Works in Cicero, Illinois, where a study was done to see the effects of lighting changes and work structure changes such as working hours and break times on worker productivity. These changes appeared to improve productivity, leading the researchers to conclude that the intervention worked. However, subsequent work suggested this was a very general effect that could be produced by almost any change to working practices.

Most therapists would regard such impacts on their clients as entirely valid outcomes of the intervention – boosting functional communication skills and confidence are key goals of any intervention package. But it is important, nevertheless, to know whether the particular intervention adopted has a specific effect. Many individualised SLT interventions with both adults and children involve complicated processes of assessment and goal-setting with the aim of achieving specific improvements. If these are actually ineffective, and the same could be achieved by simply being warm, supportive and encouraging, then we should not be spending time on them.

Generalised beneficial effects of being singled out for intervention can also operate via teachers. In a famous experiment, Rosenthal and Jackson (1963) provided teachers with arbitrary information about which children in their class were likely to be 'academic bloomers'. They subsequently showed that the children so designated obtained higher scores at a later point, even though they had started out no different from other children. In fact, despite its fame, the Rosenthal and Jackson result is based on data that, by modern standards, are far from compelling, and it is not clear just how big an effect can be produced by raising teacher expectations. Nevertheless, the study provides strong motivation for researchers to be cautious about how they introduce interventions, especially in school settings. If one class is singled out to receive a special new intervention, it is likely that the teachers will be more engaged and enthusiastic than those who remain in other classes where it is 'business as usual'. We may then be misled into thinking that the intervention is effective, when in fact it is the boost to teacher engagement and motivation that has an impact.

Study participants themselves may also be influenced by general positive impact of being in an intervention group – and indeed it has been argued that there can be an opposite effect – a nocebo effect – for those who know that they are in a control group, while others are receiving intervention. This is one reason why some studies are conducted as 'double blind' trials – meaning that neither the participant nor the experimenter knows who is in which intervention group. But this is much easier to achieve when

the intervention is a pill (when placebo pills can be designed to look the same as active pills) than when it involves communicative interaction between therapist and client. Consequently, in most studies in this field, those receiving intervention will be responding not just to the specific ingredients of the intervention, but also to any general beneficial effects of the therapeutic situation.

4.2 Identifying specific intervention effects by measures of mechanism

There are two approaches that can be taken to disentangle nonspecific effects from specific impact of a particular intervention. First, we can include an active control group who get an equivalent amount of therapeutic attention, but directed towards different goals. I will discuss this further in Chapter 5, which focuses on different approaches to control groups. The second is to include specially selected measures designed to clarify and highlight the active ingredients of the intervention. I will refer to these as 'mechanism' measures, to distinguish them from outcome measures.

Let's take the example of parent-based intervention with a late-talking toddler. In extended milieu therapy, the therapist encourages the parent to change their style of communication in naturalistic contexts in specific ways (see table x). The ultimate goal of the intervention is to enhance the child's language, but the mechanism is via changes in the parent's communicative style. If we were to find that the child's language improved relative to an untreated control group but there was no change in parental communication (our mechanism measure), then this would suggest we were seeing some general impact of the therapeutic contact, rather than the intended effect of the intervention. (Jonathan Green e.g.)

To take another example, the theory behind the computerised Fast Forword intervention maintains that children's language disorders are due to problems in auditory processing that lead them to be poor at distinguishing parts of the speech signal that are brief, non-salient or rapidly changing. The intervention involves exercises designed to improve auditory discrimination of certain types of sounds, with the expectation that improved discrimination will lead to improved general language function. If, however, we were to see improved language without the corresponding change in auditory discrimination (a mechanism measure), this would suggest that the active ingredient in the treatment is not the one proposed by the theory.

Note that in both these cases it is possible that we might see changes in our mechanism measure, without corresponding improvement in language. Thus we could see the desired changes in parental communicative style with extended milieu therapy, or improved auditory discrimination with FFW, but little change in the primary outcome measure. This would support the theoretical explanation of the intervention, while nevertheless indicating it was not effective. The approach might then either be abandoned or modified – it could be that children would need longer exposure, for instance, to produce a clear effect.

The most compelling result, however, is when there is a clear difference between an intervention and a treated control group in both the mechanism measure and the outcome measure, with the two being positively related within the group. This would look like good evidence for a specific intervention effect that was not just due to placebo.

It is not always easy to identify a 'mechanism' measure – this will depend on the nature of the intervention and how specific its goals are. For some highly specific therapies – e.g. a phonological therapy aimed at changing a specific phonological process, or a grammatical intervention that trains production of particular morphological forms, the 'mechanism' measure might be similar to the kind of 'near transfer' outcome measure that was discussed in chapter 2 – i.e., a measure of change in performance on the particular skill that is targeted. As noted above, we might want to use a broader assessment for our main outcome measure, to indicate how far there is generalisation beyond the specific skills that have been targeted.

Chapter 5

Controlling unwanted effects with a control group

The idea of a control group has already come up in chapters 3 and 4, where it has been noted that this provides one way of estimating how far an improvement from baseline to post-treatment is really due to the intervention. The logic is straightforward: if you have two groups – group A who has the intervention, and group B who does everything that group A does except for the intervention, then it should be possible to isolate the specific effect of the intervention. In this chapter I will discuss the different kinds of control group that can be used. But first, it is important to address a question of ethics.

5.1 Is it ethical to include a control group?

The argument against a control group goes something like this: we think our intervention works, but we need to do a study to establish just how well it works. It would be unethical to withhold the intervention from one group of people because they would then be deprived of a potentially helpful treatment. In fact, clinical trials developed blinding of patients, clinicians, and even the statistical analyst (single, double and triple blind studies) to allocation of treatment and placebo. This avoided the bias in conditions where patients were in an advanced stages of their condition and physicians felt morally obliged to help those most in need.

This argument has as many holes in it as a Swiss cheese, and here's why. Do you already know if the treatment works? If the answer is Yes, then why are you doing a study? Presumably it is because you think and hope it works, but don't actually have evidence that it does. So you don't know and you need to find out. It is also possible to allocate those to treatment who are likely to get the largest benefit by utilizing the regression discontinuity design. We will discuss this in later chapters, but briefly we must mention the disadvantages of this design over a randomized design. Firstly, we need substantially more observations and/or individuals for robust statistical evidence. Secondly, the relationship between treatment and outcome must be correctly modelled, and lastly, presence of other variables that might confound the result, so must be controlled.

Unfortunately, the kinds of systematic bias discussed in Chapter 3 often lead practitioners to have an unwavering conviction that their intervention is effective: they see improvement in the clients they work with, and assume it is because of what they have done. The harsh reality is that when properly controlled studies are conducted, it is often found that the intervention is not effective and that the improvement is just as great in controls as in the intervention group. There are numerous examples of this in mainstream medicine, as well as in speech and language therapy. Box x gives some examples.

- FFW
- Arrowsmith
- Cogmed

5.1.1 Possible adverse effects of intervention

Worse still, some interventions do more harm than good. The need to evaluate adverse impacts of intervention is well-recognised in mainstream medicine, where it is easy to see the increased risks of morbidity or mortality, or unpleasant side effects. In clinical trials, a new treatment or medication will pass through typically five stages or phases of trials. Each stage increases the number of inidividuals, and the early trials are specifically designed to test safety and effcicacy, beginning with very small doses. It should also be noted that prior to human trials, all treatments will undergo substantial lab based evaulation and sometimes animal testing via rodent or primate modelling.

In speech and language therapy it is typically assumed that interventions won't do any harm – and it is true that types of adverse effect seen in medicine are not likely. Nevertheless, any intervention creates opportunity costs – the time (and sometimes money) that are spent in intervention can't be spent doing other things. I first became sensitive to this issue when trying to persuade head teachers to get involved in a computerised intervention I had devised: they were keen to take part, but then discussed with me what activities children should miss if they were to spend 15 minutes every day on my computerised tasks for a period of six weeks. Should we ask children to forfeit their breaks, or to skip non-academic activities such as swimming? This would hardly be popular with the children, we felt. Or should they take time out of lessons in English or Maths? As discussed in Box 1 (FFW and Arrowsmith and Cogmed) some computerised programmes involve a substantial investment of time.

When we move to consider interventions outside the school setting, there are still time costs involved: an aphasic adult with poor mobility or a harassed mother with three small children who is reliant on public transport may have considerable difficulty making it to clinic appointments. You may reply that, even if the intervention is ineffective, the clients are glad to have the attention and professional interest of a therapist. But if that is something that is of importance to them, then that needs to be included in your outcome measures – i.e. you need to demonstrate it is so, rather than just assuming this is so.

For parent training programmes, there is an additional risk of making parents worry that they are inadequate, and that their child's difficulties are all their fault. Clearly, competent professionals will be at pains to counteract that perception, but it is an easy mindset to fall into. (Box on media reaction to Jonathan Green study?).

Finally, the person who is the focus of intervention may feel embarrassed, stressed or stigmatised by being singled out for special treatment – this is most likely to be a potential problem for children who are pulled out of regular classes for intervention sessions. Here again, a good professional will recognise that risk and do their best to address the concerns, but this needs to be thought about, rather than blithely assuming that because the intention of the therapist is good and helpful, their impact has to be beneficial.

All these potentially adverse impacts, including the large investments of time associated with the kinds of intervention shown in Box x, will generally be seen as a price well worth paying if the intervention is effective. At the end of the day, a decision whether to intervene or not should involve an appraisal of costs and benefits. Unfortunately, it is all too common to completely ignore the costs and to adopt a rosy, but unevidenced, view of benefits.

5.1.2 Uncontrolled studies are unethical

As explained in Chapter 3, a study that just compares people on a language measure before and after intervention is generally uninterpretable, because there are numerous factors that could be responsible for change. Having a control group is not the only way forward: in Chapter x, I discuss other types of research design that may be more appropriate for specific contexts. But in most situations where the aim is to test the effectiveness of an intervention for a particular condition, a study with a control group will be the best way to get useful information. Indeed, doing a study without controls is unethical, because you end up investing the time of clients and professionals, and research funds, doing a study that cannot answer the question of whether the intervention is effective.

5.2 Treated vs untreated controls

As noted above, a common design is to compare group A, who receive intervention, with group B who are treated just the same but without any intervention. If you see no effect of intervention using this design, and the study is big enough to give adequate statistical power (see Chapter 12), then you can be pretty certain that the intervention is not effective. But if you do see a reliable improvement in group A, over and above and change in group B, can you be certain it was the intervention that made the difference?

There are two things to be concerned about here. First, in the previous chapter I discussed the kinds of very general impact that can be caused just by being in an intervention group. These work in the opposite way to the possible adverse effects discussed above: although some people may be embarrassed or stressed by being a target of intervention, others may be reassured or made more confident. A good therapist will do more than just administer an intervention in a standard fashion: they will form a relationship with their client which can itself be therapeutic. This, however, creates a confound if the control group, B, just has 'business as usual' because it means we are really comparing group A, who gets a specific intervention plus a therapeutic relationship, with group B, who gets nothing.

- social vs language?
- Dental care?
- Gillam study?

A creative way of dealing with this confound is to ensure that group B does also get some kind of intervention, but different from the one that is the focus of the study. For instance, (Snowling study) compared children who had a parent-reading intervention with a group who had a parent-based intervention designed to improve motor skills. Box X shows other examples (social vs language? Dental care? – Gillam study?) It is important to choose a comparison intervention that is not expected to have an impact on the main outcome measure – otherwise you could find yourself in a situation where everyone gets better but groups A and B don't differ, and you are left uncertain whether both interventions worked or neither did.

The use of a contrasting intervention overcomes another possible concern about untreated controls, which is those in group B may actually do worse because they know they are missing out on an intervention that might be able to help them (see Chapter 4).

Table 5.1: Treatment Allocation Matrix

	Period 1	Period 2
Group 1	A	B
Group 2	B	A

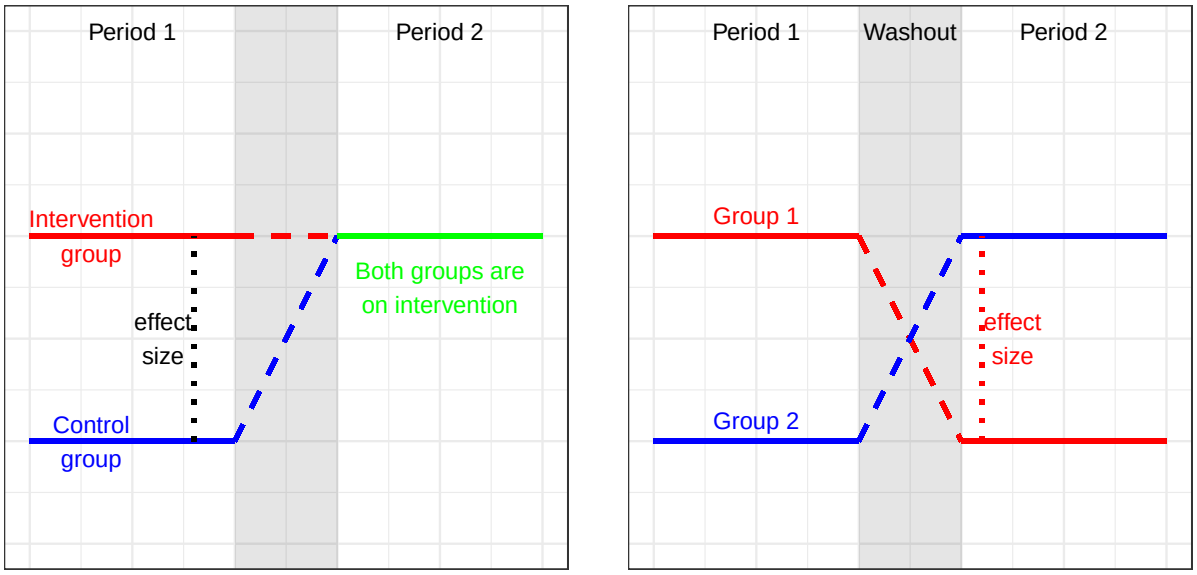


Figure 5.1: plotting showing a crossover design RCT

5.3 Wait-list controls and cross-over designs

Another popular approach to controls is to have a two-stage study. Stage 1 is a classic controlled study where a treated group A is compared with an untreated group B. After the intervention phase, group B is then offered the intervention. At this point, group A may continue with the intervention, or intervention may be withdrawn.

Possible outcomes of such a study are shown in Figure x. The pattern seen in panel 1 is unlikely for an intervention that is intended to produce long-lasting effects – in SLT practice, it would not be expected that skills would decline after an effective intervention had been withdrawn. (more needed on this)

5.3.1 Cross-over design

The benefit of the crossover design is that a potentially more accurate evaluation of intervention comparison is achieved as we compare individuals to themselves rather than controls that are different individuals. A secondary benefit is that crossover designs typically require fewer individuals as a separate control group is not necessary. A crossover deign is split into three phases, an initial phase where two groups are randomized (exactly the same as a parallel group design) to intervention and control. Once the first phase has ellapse, the individuals enter a washout phase, this is important to allow any intervention effect to be removed before the groups are switched and phase two initiates, so that both groups have received intervention. The assessment of the intervention looks at the differences between phases 1 and 3 by group. We hope to see no group difference but a significant phase difference.

Chapter 6

Observational studies

Janice, a SLT, has recently started using a parent-based approach with late-talking 2-year-olds in her practice. Parents are encouraged to make video recordings of interactions with their child, which are then analysed with the therapist, who notes ways of making the interaction more contingent on the child's interests and communicative attempts. She wants to evaluate what she is doing. A colleague of hers, Anna, who has a similar caseload, is sceptical about whether Janice's approach is cost-effective. Anna uses a watchful waiting approach with children this young. Janice and Anna agree to do a study using the same pre- and post-intervention assessments so that they can evaluate the impact of Janice's approach.

Stephen notes that some aphasic patients referred to his clinic talk glowingly about a commercially available 'brain-based learning' programme, MyLangBooster. He wants to know whether he should be recommending this programme to other clients, so he carries out a comparison of patients who used MyLangBooster and those who did not.

Dorothy, a researcher, had a grant for a study predicting outcomes of 4-year-olds who had poor language skills, some of whom had received intervention with SLT services. She saw the same group of 83 children at ages 4, 4.5 and 5.5 years, and noted that their outcomes were very variable. Some children had caught up with their peer group by the final assessment, whereas others had persistent problems. When she presented the results at a conference, a person in the audience suggested that she should do an analysis to see if the amount of intervention was a factor influencing outcomes.

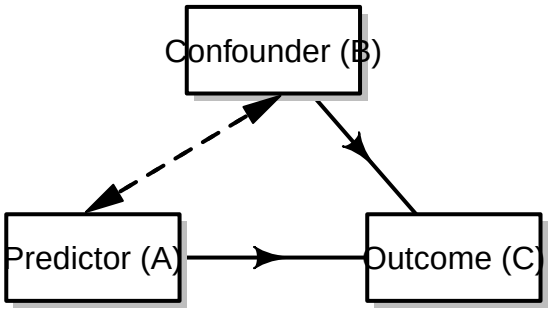
All of these vignettes illustrate observational studies: ones where we use naturalistic data rather than exerting any experimental control over who gets an intervention. The last one, as you may have guessed, is a true story – my own experience with a longitudinal study by Bishop and Edmundson (1987). This gives a very clear illustration of the perils of observational studies.

When I looked at the relationship between intervention and outcome in my sample, I found results that were deeply depressing to SLTs: the children who had the most intervention had the worst outcomes. Did that mean that intervention was actually harming children? Well, it's possible – as noted in chapter 5, it is dangerous to assume that all effects of intervention are benign. But there was a much simpler explanation for this topsy-turvy result: the children who received intervention were different from those who didn't – in general they had more severe problems to start with. This makes sense: if a child is referred to SLT services, then the therapist makes a judgement about which ones to intervene with, and these are likely to be different from those who are discharged or simply reviewed after a few months. If, as appears to have been the case in my study in the 1980s, therapists work most with the more serious cases, then intervention may appear ineffective. On the other hand, a contemporary concern in the UK is that therapists are being encouraged to focus their efforts on children who can be discharged after a short block of treatment, which means they prefer to work with children with milder problems. This will create the opposite impression – therapy which is actually ineffective may appear effective. The basic problem is the same: treated and untreated groups are not comparable, and so comparing them will give misleading results.

Stephen's study is equally problematic. Here we are comparing a self-selected group of patients with his regular caseload. Those who tried MyLangBooster may be more motivated to improve than other patients. They may have more money, so they can afford to pay for the programme. Or they may be more desperate – having tried other interventions that failed to make a difference. Furthermore, Stephen may only hear from those who felt they improved, and be unaware of other patients who tried it but then dropped out because they obtained disappointing results. It is almost never a good idea to base an evaluation of an intervention on a study of self-selected enthusiasts. There are just too many potential confounds that could cause bias.

What about the case of Janice and Anna? This may seem less problematic, since the two therapists have similar caseloads, and the decision about therapy is based on therapist preference rather than child characteristics. Here again, though, the comparison has the potential to mislead. If baseline and post-intervention assessments are done using the same measures, then it is at least possible to check if the children in the two practices are similar at the outset. But there would still be concerns about possible differences between the therapists and their practices that might be influencing results. Maybe Anna rejects parent-based intervention because she knows that most of the parents in her practice have full-time jobs and would not be willing or able to devote time to attending sessions to analyse videos. Maybe Janice is an exceptionally skilled therapist who would obtain good outcomes with children regardless of what she did. Perhaps her enthusiasm for a parent-based approach contrasts with Anna's more downbeat attitude, and this has an influence on parent and/or child. In sum, there is scope for all the non-specific treatment effects discussed in Chapter 4 to exert an impact. If Janice finds better outcomes than Anna, despite doing their best to ensure that the participating children and parents from their practices are similar, then it is reasonable to say that this is useful information that would provide justification doing an experimental study (see Chapter x). But it is not conclusive and cannot substitute for the kind of experimental study discussed in the next chapter.

Things that may affect outcome and that differ between intervention and control groups are known as **confounders**.



Consider possible confounders in the following examples: Does long-term use of hormone replacement therapy carry risks or benefits? Does excessive use of computer games in teenagers cause social isolation? Will your life be extended if you eat more cabbage? Here are just a few possibilities: Woman who decide to continue to use HRT may have more severe menopausal symptoms. Excessive use of computer games may be a consequence rather than a cause of lack of social engagement, and those who eat cabbage may adopt healthier lifestyles than

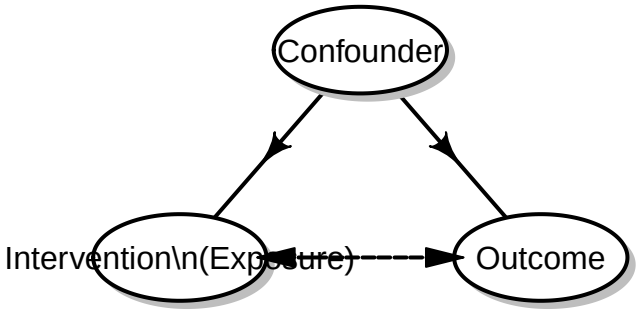


Figure 6.1: Alternative interpretation of confounder path diagram

those who don't.

Most reports in the media are based on observational rather than experimental data. In some cases, it is hard to see how other types of data could be collected: are we really going to succeed in allocating adolescents to a regime of non-stop computer games, or to force people to start eating cabbage? Indeed, some associations that are now well-established, such as the link between cigarette smoking and cancer, were discovered from observational data, and could not be studied any other way. But where the interest is in interventions administered by a therapist, then it should not be necessary to rely on observational studies, and, as we have shown, to do so can lead to flawed conclusions.

6.1 Class exercise

Find a newspaper report of a factor that is reported to be a risk or benefit to health. Is it based on an experimental or observational study? Can you identify potential sources of bias?

6.2 Observational study designs

The key difference between observational studies and other types discussed in this book are that the researcher(s) do not intervene at any stage, and simply observe whether a condition develops, and record any risk factors or particular behaviour.(cite: Sedgewick 2012). The difference between the designs discussed below is either the point at which the intervention or outcome are established.

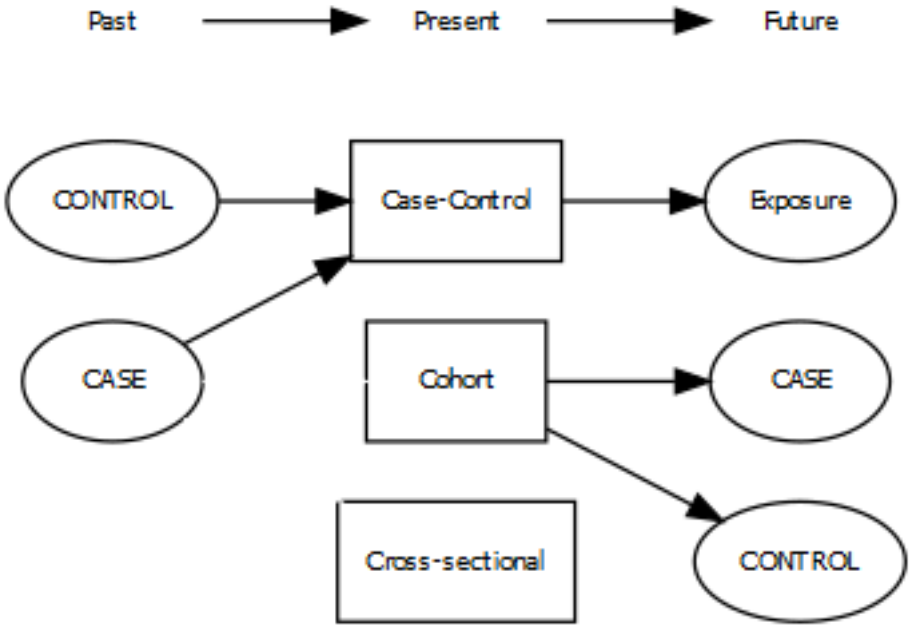


Figure 6.2: Figure 6.3: time course of common observational studies

6.2.1 Cohort studies

A Cohort study follows a group of individuals who do not have the condition or disorder of interest at the time that they enter the study, but they are selected from an at-risk population. The cohort is followed up, observing those that develop the condition or disorder, and those that do not (forming a natural control group, sometimes called an internal control, see Mann, 2003). Within the cohort design, there are different variations, single cohort as above, or two cohort design . For example, a two cohort design would collect data on two cohorts of individuals (both groups do not have the condition at the start of the study). One cohort is assigned to intervention, the other cohort is a control so would be assigned a sham intervention perhaps. Both cohorts would be followed up and the observation of who developed the condition and who did not is recorded for both cohorts.

6.2.2 Case-Control studies

Typically, the case-control study looks at retrospective information on a group of individuals, both a control group (those without particular condition or disorder) and cases (those with a known condition or disorder). Individuals are chosen purely on their status of condition or disorder of interest. These individuals are then asked about previous exposure to potential risk factors, for example, if we were looking at a reading disorder in school-aged children, we might want to record information on family history of the condition, or exposure to books in the home environment, or whether the children and their parent or guardian co-read at home at a pre-school age. The choice of controls participants is key as they must form a representative sample of individuals that are within the at-risk group population but do not have the condition or disorder of interest.

6.2.3 Cross-Sectional Studies

The cross-sectional study is used when only one time point is recorded and its primary purpose is typically to establish prevalance of conditions or disorders. In this design, we differentiate from the others designs as the intervention and outcome are determined simultaneously (cite: Carlson & Morrrison, 2009).

6.3 STROBE guidelines

The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies was developed to try to standardize the key details reported in observational studies in epidemiology (with a focus on the medical field). These guidelines have been adopted to counter poor reporting, the authors of those guidlines state the purpose is not to be a rigid template or assessment tool, but

> issues such as confounding, bias, and generalisability could become more transparent, which might help temper the over-enthusiastic reporting of new findings in the scientific community and popular media, and improve the methodology of studies in the long term. Better reporting may also help to have more informed decisions about when new studies are needed, and what they should address.
>
> - Von Elm et al (2007)

The Strobe guidelines themselves consist of 22 items that are a necessary requirement to report to ensure some degree of transparency to the research, and ensure a good standard of reporting quality. We restate the STROBE guidelines from Von Elm at al. (2007) in **??**

	Item number	Recommendation
TITLE and ABSTRACT	1	▪ Indicate the study's design with a commonly used term in the title or the abstract ▪ Provide in the abstract an informative and balanced summary of what was done and what was found
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported
Objectives	3	State specific objectives, including any prespecified hypotheses
Study design	4	Present key elements of study design early in the paper
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection
Participants	6	▪ Cohort study—Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up Case-control study—Give the eligibility criteria, and the sources and methods of case ascertainment and control selection. Give the rationale for the choice of cases and controls Cross-sectional study—Give the eligibility criteria, and the sources and methods of selection of participants ▪ Cohort study—For matched studies, give matching criteria and number of exposed and unexposed Case-control study—For matched studies, give matching criteria and the number of controls per case

	Item number	Recommendation
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable
Data sources/ measurement	8	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group
Bias	9	Describe any efforts to address potential sources of bias
Study size	10	Explain how the study size was arrived at
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen, and why
Statistical methods	12	▪ Describe all statistical methods, including those used to control for confounding ▪ Describe any methods used to examine subgroups and interactions ▪ Explain how missing data were addressed ▪ Cohort study—If applicable, explain how loss to follow-up was addressed Case-control study — If applicable, explain how matching of cases and controls was addressed Cross-sectional study — If applicable, describe analytical methods taking account of sampling strategy ▪ Describe any sensitivity analyses
Participants	13	▪ Report the numbers of individuals at each stage of the study—e.g., numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed ▪ Give reasons for non-participation at each stage ▪ Consider use of a flow diagram

	Item number	Recommendation
Descriptive data	14	▪ Give characteristics of study participants (e.g., demographic, clinical, social) and information on exposures and potential confounders ▪ Indicate the number of participants with missing data for each variable of interest ▪ Cohort study—Summarise follow-up time (e.g., average and total amount)
Outcome data	15	Cohort study—Report numbers of outcome events or summary measures over time Case-control study—Report numbers in each exposure category, or summary measures of exposure Cross-sectional study—Report numbers of outcome events or summary measures
Main results	16	▪ Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (e.g., 95% confidence interval). Make clear which confounders were adjusted for and why they were included ▪ Report category boundaries when continuous variables were categorized ▪ If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period
Other analyses	17	Report other analyses doneâ€"e.g., analyses of subgroups and interactions, and sensitivity analyses
Key results	18	Summarise key results with reference to study objectives
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence
Generalisability	21	Discuss the generalisability (external validity) of the study results
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based

The guidelines follow the structure of a typical research article, breaking down the required elements under subheadings that are compatable to most journals reporting observational research.

6.4 Matching

6.4.1 Propensity score

6.4.2 distance matrices

6.4.3 Optimal matching

Chapter 7

Controlling for selection bias: randomised assignment to intervention

In chapter 6, we saw how observational studies could be misleading because they are likely to contain unknown differences between intervention and control groups. Recognition of this limitation led to the development of the randomised controlled trial as a gold standard approach to the evaluation of interventions. The core idea is that once a target population has been identified for study, the experimenter allocates participants to intervention and control groups in a way that avoids any bias.

There is a quite substantial literature on methods of allocation to intervention. The main principle behind modern intervention studies is that allocation should be completely random, i.e. not predictable from any characteristics of participants. There are many approaches to treatment allocation that are designed to avoid bias but which are not random. For the time being we are ignoring the question of who does the randomisation: this will be discussed in Chapter 8. We also defer until later the question of recruitment and what to do about people who do not wish to take part in a trial. We assume we have a population of potential participants and have to decide who will be in the intervention group and who will be in the control group. Consider the following methods of allocation and note your judgement as to whether they are (a) reasonable ways of avoiding bias and (b) random.

A. Allocate people to the intervention group until the desired sample size is reached; then allocate all subsequent cases to the control group
B. Allocate the more severe cases to the intervention group and the less severe to the control group
C. Alternate between intervention and control: thus the first person on the list is an intervention case and the last one is a control
D. For each potential participant, toss a coin and allocate those with 'heads' to the intervention group and those with 'tails' to the control group
E. Create pairs of people matched on a key variable such as age, and then toss a coin to decide which will be in the intervention group and which in the control group
F. Create pairs of people matched on the baseline score on the outcome measure, and then toss a coin to decide which will be in the intervention group and which in the control group

If you have read this far, it is hoped that you will identify that B is clearly a flawed approach. Not only does it fail to equate the intervention and control groups at the outset, it also is likely to result in regression to the mean, so that greater gains will be seen for the intervention vs. the control group, for purely spurious statistical reasons. Despite its obvious flaws, this approach is sometimes seen in published literature – often taken to the extreme where an intervention group with a disorder is compared to a 'control' group with no disorder. This makes no sense at all – you are not controlling for anything unless the groups are equivalent at the outset, and so cannot evaluate the intervention this way.

Approach A is an improvement, but it is still prone to bias, because there may be systematic differences between people who are identified early in the recruitment phase of a study and later on. For instance, those who sign

up immediately may be the most enthusiastic. Approach A also creates problems for blinding, which is discussed further in Chapter 9.

Most people think that Approach C as a reasonable way to achieve equal group sizes and avoid the kinds of 'time of recruitment' issues that arise with Approach A. But this is not a random approach and would therefore be regarded as problematic by modern standards of trials design. One reason, which I will discuss at more length in Chapter 8, is that this method could allow the researcher to influence who gets into the study. Suppose the experimenter allocates aphasic patients to a study to improve word-finding according to some systematic method such as their serial order on a list, but notices that the next person destined for the intervention group has particularly profound difficulties with comprehension. It may be sorely tempting to ensure that he had an 'odd' number rather than an 'even' number and so ended up in the control group. Similar objections arise to methods such as using a person's case number or date of birth as the basis for intervention assignment: although this may be a fixed characteristic of the person and so appear to avoid bias, it raises the possibility that the experimenter could simply decide not to include someone in the study if they were destined for the intervention group and looked unpromising.

Approach D is clearly random, but it runs the risk that the two groups may be of unequal size and they may also be poorly matched on the pre-intervention measures, especially if sample sizes are small.

Approach E would seem like the logical solution: it avoids unequal sample sizes, ensures groups are comparable at the outset, yet includes randomisation. This method does, however, have some disadvantages: it can be difficult to apply if one does not have information about the full sample in advance, and can be problematic if participants drop out of the trial, so the matching is disrupted.

7.1 Units of analysis: Individuals vs clusters

In the examples given above, allocation of intervention is done at the level of the individual. There are contexts, however, where it makes sense to use larger units containing groups of people such as classrooms, schools or clinics. Research in schools, in particular, may not readily lend itself to individualised methods, because that could involve different children in the same class receiving different interventions. This can make children more aware of which group they are in, and it can also lead to 'spill-over' effects, whereby the intervention received by the first group affects other children who interact with them. It may make sense for children in classroom A to have one intervention and those in classroom B to have a control intervention. This is what happens in a clustered trial.

This method, however, raises new problems, because we now have a confound between classroom and intervention. Suppose the teacher in classroom A is a better teacher, or it just so happens that classroom B contains a higher proportion of disruptive pupils. In effect, we can no longer treat the individual as the unit of analysis. In addition, we may expect the children within a classroom to be more similar to one another than those in different classrooms, and this has an impact on the way in which the study needs to be analysed. Clustered studies typically need bigger sample sizes than non-clustered ones. We will discuss this issue further in Chapter 10.

7.2 Class exercise

Use a literature search for an intervention/disorder of interest with the term 'randomised controlled trial' or RCT, and download the article. Is the randomisation process clearly described? Is it adequate?

7.3 Randomization methods

7.3.1 Simple randomization

In our earlier examples from the start of this chapter, we saw various examples of simple randomization (A-D). This takes the idea of the "coin toss" but adds in some degree of control of other variables which might affect the result. Typically, a coin is not used as the random nature of the coin toss means that balanced numbers in each group will not be guaranteed. A randomization list is created by an individual who is unrelated to the trial. This list contains an equal number of allocations for a fixed number of participants which will have been pre-specified according to a sample size calculation. Our first issue is that the allocation list does not allow for individuals dropping out before assignment of intervention. This means that we potentially can have unbalanced groups despite following the list.

Furthermore we saw earlier, simple randomization does not suffciently control unknown variables which may exhert an effect on the outcome, i.e. overrepresentation of a particular feature in one group that influences the outcome. For example, we randomize children with a language difficultly into two groups, but we find that there are more boys than girls in one group. The intervention sees an effect as predicted, but on further inspection we find that the group difference is confounded by gender.

Simple randomization is rarely (if at all) used in practice, so other approaches have been developed to counter the drawbacks encoutered.

7.3.2 Random permuted blocks

The permutated blocks approach tries to overcome the unbalanced problem by allocating individuals into blocks. The idea is that patients are allocated to intervention or control in blocks, so that at certain equally spaced points across the trial, equal numbers are ensured. Let's see how this works using an example:

Say we have chosen out block size to be 4. So, in each block we have 2 assigned to intervention and 2 to control group. The order of assignment of each block is random, for example, ABAB, AABB, ABBA, BBAA, BAAB, and so on. This means at any point in the trial, the randomization is only ever two individuals unbalanced if the trial stops short or fails to recruit sufficient numbers of inidividuals, so the imbalance is minimal and statistical power is maintained.

However, permuted blocks has an unfortunate downside as it is relatively straighforward for the researcher to see a pattern emerge in the allocation, so the randomization can become predictable. This is problematic, if the next allocation is known to the researcher as control and the next child for allocation is a particularly severe case who could benefit from the intervention. This would rely on the researcher making a difficult moral judgement to continue with correct allocation and not deviate.

7.3.3 Unequal randomization

The approach of unequal randomization differs from the convention of 1:1 allocation ratio. A fixed ratio is decided on before allocation of individuals begins, for example, 2:1 intervention vs control allocation. Use of this method should be considered carefully as there are both pros and cons. On the positive side, we may obtain a more accurate estimate of the effects of intervention, but we are prone to a reduction in power.

> "Unequal allocation also has consequences for statistical power. For example, a 2:1 allocation ratio requires 12% more patients than a trial using 1:1 to detect the same size effect with equivalent power; a 3:1 randomization scheme requires 33% more patients."

Hey, S. P., & Kimmelman, J. (2014). The questionable use of unequal allocation in confirmatory trials. Neurology, 82(1), 77-9.

Table 7.1: Randomized allocation by stratas: Age, School

	Intervention	Control
school1, ages 5-7	23	24
school2, ages 8-9	31	31
school3, ages 10-11	28	27
school1, ages 8-9	19	20
school2, ages 10-11	25	24
school3, ages 5-7	35	35
school1, ages 10-11	22	27
school2, ages 5-7	24	23
school3, ages 8-9	31	31

Table 7.2: Randomized allocation by minimization

Prognostic factor	Intervention	Control
Sex		
Male	3	5
Female	5	3
Age band		
21-30	4	4
31-40	2	3
41-50	2	1
Risk factor		
High	4	5
Low	4	3

7.3.4 Stratification

The previously discussed randomizations have dealt with the unbalanced groups but we are still in need of a way to counter the effect of known or unknown variables on the outcome. Stratification is the first method to permit some control of these variables by splitting the allocation according to those variables. For example, suppose we have an intervention for school aged children and we have a exclusion criteria that they must be between the ages of 5 and 11. We also provide the intervention at three different schools, one inner city school in a low SES area, one public school requiring fees, and a small remote village school. We would anticipate that there might be significant performance differences in age and potentially differences between schools, so we must allow for this when we randomize children to intervention and control groups. The scheme would follow the table below,

7.3.5 Minimization

An alternative to stratification called minimization is also very popular in clinical RCTs. Its popularity has grown in recent years as stratifcation becomes weaker when the number of strata increases and is particularly susceptable when used in smaller trials. Minimization was first proposed by Pocock and Simon (ref) and is referred to as a dynamic or adaptive randomization. Each participant's allocation is dependant on the characteristics and allocation of participants already allocated to groups. The idea is to minimise the imbalance by considering a range of factors for each allocation. This prevents the allocation of particular sub groups into one allocation which may happen in any of the aforementioned methods. To demostrate this method, we restate the example from Scott et al. (2002), using table 7.2:

> If we now consider allocation of the 17th participant who has factors: Male, 31-40, and High. **Using the Pocock and Simon's method**

If allocated to intervention group, total imbalance is

$$|(3+1)-5|+|(2+1)-3|+|(4+1)-5|=1$$

. If allocated to control group, total imbalance is

$$|3-(5+1)|+|2-(3+1)|+|4-(5+1)|=7$$

. Patient allocated to the group that would lead to less overall imbalance. Therefore 17th participant allocated to intervention group because

$$1<7$$

.

Chapter 8

The experimenter as a source of bias

We have discussed numerous sources of bias that can occur in a study: systematic changes over time, use of inadequate measures, and inappropriate intervention groups. We now come to consider one further important factor: the experimenter.

The role of the experimenter was already alluded to in Chapter 7, where it was noted that some kinds of randomisation are problematic because they give the experimenter wiggle room to omit or include participants. You might imagine that no honest person would distort their study this way, but it is all too easy to provide a post hoc rationalisation for such decision. I realised this problem when I first tried to do a randomised trial myself. I was visiting a school to where we had recruited children for the study, and I was chatting to the children in their classroom. One little boy was overactive and seemed unable to concentrate for more than a minute or two on anything, yet he had met our criteria for study inclusion. I found myself fervently hoping he would not be in the intervention group! Fortunately, I had no control over this, and indeed, to this day I don't know if he was.

8.1 Allocation concealment

This brings us to the topic of masking, or allocation concealment about intervention status from the experimenter. Traditionally, the term 'blinding' has been used, but this is offensive to some visually impaired people, so I will avoid it here. Nevertheless, you may hear the term in the context of the 'double blind trial'. This refers to the situation when the neither the experimenter nor the participants are aware of who is in the active or the control group. As noted in chapter 4, it is seldom feasible to keep participants unaware of whether they are in an active intervention or control group, though use of active controls, as described in Chapter x, may allow for this.

In SLT practice, where resources are limited and the same person may be conducting the study and administering the intervention, particular care needs to be given to masking. A competent third party should be recruited to be responsible for allocation to intervention groups, but this is not all. It is also important to take steps to avoid experimenter bias in administration of baseline and outcome assessments.

8.2 The importance of masking for assessments

As discussed in Chapter 2, some assessments are more objective than others: it is relatively easy to be swayed by one's own desires and expectation when making a global rating of a person's communicative competence on a 4-point scale, much less so when administering a multiple choice comprehension test, where the participant selects a picture to match a named word. Nevertheless, there is ample evidence that, even with relatively objective tasks, experimenter bias can creep in. (Egs of people measuring wt/ht).

Perhaps the most sobering example comes, not from an intervention study, but from data on a phonics screening test administered by teachers in UK schools in the years from xxxx to xxxx. The score is supposed to just indicate the number of items on a standard list of words and nonwords that a child reads accurately. Although there is some leeway in judging whether a nonword is correctly pronounced, this should not be expected to exert a big effect on final scores, given that teachers are provided with a list of acceptable pronunciations. The results, however, which are published annually on a government website, show clear evidence of scores being nudged up for some cases. We would expect a normal distribution of scores, but instead there is a sudden dip just below the pass mark and a corresponding excess of cases just about the pass mark. Since teachers were aware of the pass mark, they would have been able to nudge up the scores of children who were just one or two points below, and the distribution of scores is clear evidence that this happened.

We don't usually think of teachers as dishonest or fraudulent. Some were opposed in principle to the phonics check and may have felt justified in not taking it seriously, but I suspect that most were doing their sincere best, but just did not think a slight tweak of a child's results would matter It's likely that many did not like the idea of categorising children as 'failing' at such a young age. And some may have been concerned that their personal reputation or that of their school might be damaged if too many children 'failed'. The possible reasons for 'nudging' are many, and cannot be established post hoc, but the point I want to stress is that this kind of behaviour is very common, not typically done in a deliberately dishonest fashion, but is something that will happen if people are motivated to get one result or another.

In medical trials, the large amount of evidence for experimenter bias on measures has led to a general consensus that it is vital that outcome assessments much be done by someone who is unaware of whether the participants was in the intervention or control group. This is likely to add to the cost of conducting a trial, but it gives security that no nudging has taken place.

8.3 Conflict of interest

Many medical journals, and increasingly journals in other fields, require author to declare conflicts of interest, and a statement to this effect is included with the published paper. Lo and Field (2009) define conflict of interest (COI) as: "a set of circumstances that creates a risk that professional judgement or actions regarding a primary interest will be unduly influenced by a secondary interest". Typically, people think of COI as involving money. Someone who is marketing a computerised intervention, for instance, will be motivated to show it is effective – their business and livelihood would collapse if it was widely known to be useless. But COI – also sometimes referred to as 'competing interests' – extend beyond the realms of finance. - A researcher's reputation may depend heavily on their invention of an intervention approach - A therapist may be aware of threats to cut government-funded services unless intervention is shown to be effective.

For those in vocational professions such as therapy or teaching, relatively poor working conditions may be endured in return for a sense of doing good. An intervention study with null results can be hard to accept if it means that the value of one's daily work is challenged.

In effect, most people involved in intervention studies want to show a positive effect for one reason or another. It's not generally possible to avoid all conflict of interest, but the important thing is to recognise experimenter bias as the rule rather than the exception, identify possible threats this poses to study validity, and take stringent steps to counteract these. I discussed above the ways in which results on outcome measures may be nudged up or down, often without any conscious attempt to mislead. But the impact of experimenter bias can occur at all stages of an intervention study:

- At the stage of study design, the researcher may argue against including a control group – claiming ethical objections – because they are aware that it is much easier to show apparent intervention effects when there are no controls (see Chapter x)
- In a controlled study, when allocating participants to intervention or control groups, the researcher may change inclusion criteria as the study progresses
- Allocation to intervention or control groups may be under the control of a researcher who does not adopt truly random methods, and so can determine who is in which group. Chapter X explains how randomisation

can overcome this.

- As noted above, if the person doing baseline and outcome assessments knows which intervention group a participant is in, then scores may be nudged. This can be avoided by having assessments done by someone who is unaware of who got the intervention
- When it comes to analysing the data, the researcher may decide on which variables to report, depending on the results. I discuss this problem and how to counteract it in Chapter x.
- If the pattern of results does not show that the intervention is effective, then further checks and alternative analyses are conducted, whereas positive results are accepted without additional scrutiny
- If the trial gives disappointing results, then the decision may be made not to report them. See Chapter X for discussion of this problem and suggestions for avoiding it.

The important point to recognise is that being a good scientist often conflicts with our natural human tendencies. A good scientist is always objective, and interpretation of results is not swayed by personal likes and dislikes. On getting a positive intervention result, a good scientist will immediately ask: "Were there biases in my study that could have contributed to this finding?" – and indeed will not take offence if other scientists identify such factors. We need, of course, arguments in the opposite direction: a failure to see an intervention effect doesn't necessarily mean the intervention did not work – there could be aspects of study design that hide true effects, including too small a sample (see Chapter x). But I find that when I attend conferences where people discuss intervention studies, the question session is always dominated by people presenting arguments about why a null result may be misleading, but it is much rarer to hear people questioning the validity of a positive result. It is a human tendency to accept information that fits with our desires and prejudices (in this case, that intervention is effective) and to reject contrary evidence. It also goes against our human nature to be critical of an intervention study conducted by a well-intentioned person who has put in a great deal of time and effort. But at the end of the day it is the quality of the evidence, rather than the feelings of researchers, that must take priority. Currently, we still know rather little about how best to help children and adults who struggle with speech, language and communication. We will only change that if we take a rigorous and evidence-based approach to evaluating evidence.

8.4 Class exercise

Once again, you need to find a published intervention study – this could be one you selected for a previous exercise, or a new one. - Does the published paper include a 'conflict of interest' or 'competing interests' statement? - List the possible factors that might lead the experimenter to be biased in favour of finding a positive result - Consider the list of stages in the research process where experimenter bias could affect results: How have the researchers counteracted these?

Chapter 9

Further potential for bias: who drops out and who volunteers?

Marie is evaluating a phonological awareness training package with a group of 'at risk' five-year-olds. She has adopted a randomised controlled design, with an active control group who get extra help with maths. She recruits 30 children to each group and runs the intervention for six weeks. However, two weeks in to the study, three children in the phonological awareness group and one from the control group drop out, and she learns that another child from the control group has been taken by his mother for intensive private speech and language therapy which includes phonological awareness training. Marie is left unsure what to do? Should she exclude these children from her study?

9.1 Dealing with dropouts: Intention to treat analysis

Every information sheet for an intervention study emphasises that participants are free to drop out at any time, without giving a reason, with no adverse consequences. This is as it should be: it would not be ethical to compel people to take part in research, and it would be foolish to try. Nevertheless, every researcher's heart sinks when this happens because it can really mess up a trial and make it harder to show an intervention effect.

If Marie were to consult a statistician, she would get an answer that might surprise you. She should not exclude participants who drop out or start other interventions. If people have withdrawn from the study, then their data cannot be included, but if there are children who are willing to be seen for the outcome assessments, their data should be included, even if they did not do the intervention they were assigned as planned. At first glance, this seems crazy. We want to evaluate the intervention: surely we should not include people who don't get the full intervention! Even worse, the control group are supposed to contain people who didn't get the intervention – if we include some who obtained it by other means, this reduces the chances of seeing an effect.

The problems posed by drop-outs may be illustrated with an extreme example. Suppose Bridget Jones volunteers for a study of a miracle cure for fatness that involves giving up regular food for six weeks and consuming only a special banana milkshake that provides 500 calories per day. The control group consists of people who want to lose weight but are just given written materials explaining the calorie content of food. At the end of the six weeks, only 5 of the 50 original milkshake group are sticking to the diet, but they have all lost impressive amounts of weight compared to the control group, whose weight has gone down marginally. Bridget, together with 90% of the intervention group, found the diet impossible to stick to, got so hungry that she binged on chocolate, and actually gained weight. Is it reasonable to conclude that the milkshake diet worked? It depends on what question you are asking. If the question is 'does extreme calorie restriction lead to weight loss?' the answer would appear to be yes. But that's not really what the study set out to do. It was designed to ask whether this specific intervention was effective for the kind of people enrolled in the study. Clearly, it wasn't – and it would be a bad idea to roll this out to larger samples of people.

In more realistic scenarios for SLTs, interventions are not likely to be so unacceptable to large numbers, but a key point is that drop-outs are seldom random. There are two opposite reasons why someone might give up on an intervention: they may just find it too hard to continue – and this could be because of difficulties in finding the time as much as difficulties with the intervention itself – or they might decide they are doing so well that they no longer need it. Either way, the group left behind after drop out is different from the one we started with.

It's instructive to look at the results from the little study I did using computerised tasks to train comprehension and spelling in children (Chapter x). Teachers and teaching assistants did the exercises with the children, three of whom dropped out. When I looked at the data, I foud that the drop-outs had particularly low comprehension scores at the baseline. It is likely that these children just found the exercises too hard and became dispirited and unwilling to continue. Importantly, they were not a random subset of children. If I had excluded them from the analysis, I would no longer be comparing randomised groups who had been comparable at baseline – I would be comparing an intervention group that now omitted those with the worst problems with a control group that had no such omissions.

Nevertheless, if we only do an 'intention-to-treat' analysis, then we lose an opportunity to learn more about the intervention. I would always complement it with an analysis of the drop-outs. Did they look different at baseline? Do we know why they dropped out? And what were their outcomes? For my comprehension study, this analysis gave a salutary message. The drop-outs showed substantial improvement on the outcome comprehension test, providing a clear example of how scores can change with practice on the test and regression to the mean (see Chapter x).

9.2 Who volunteers for research?

Many people who might benefit from interventions never take part in research evaluations. It has been noted that in psychological research, there is a bias for participants to be WEIRD, that is Western, educated, and from industrialized, rich, and democratic countries. I am not aware of specific studies on participants in SLT intervention studies, but it is likely that similar biases operate. It is important to know about the demographics of participants, as this may affect the extent to which results will be generalizable to the population at large. Researchers should report on the distribution of socio-economic, educational and ethnic background of those taking part in intervention studies.

If the aim is to evaluate intervention targeted at disadvantaged groups, then it is important to think hard about how to optimise recruitment. Positive moves might include having a researcher or other intermediary who comes from a similar background to the target population, who will know the best ways to communicate and gain their confidence. Running a focus group may give insights into barriers to participation, which may be practical (finding time, or costs for transport of childcare) or social (feeling uncomfortable around people from a different social or ethnic background). Small things can make a big impact: one study with disadvantaged families that involved parent-training ended with parents being given a certificate on completing the training: the researchers said that this was appreciated by many parents, most of whom had never obtained any formal qualifications. Some of them went on to help recruit and train the next parent group.

9.3 Class exercise

Download a copy of the CONSORT flow diagram: http://www.consort-statement.org/consort-statement/flow-diagram Use of this kind of diagram has become standard in medical trials, because it helps clarify the issues covered in this chapter. Find an intervention study in the published literature which does not include a flow diagram, and see how much of the information in the flow chart can be completed by reading the Methods section.

Chapter 10

Putting it all together: the randomised controlled trial

The randomised controlled trial (RCT) is regarded by many as the gold standard method for evaluating interventions. In later chapters I will discuss some of the limitations of this approach that may make it less than ideal for evaluating SLT interventions. But in this chapter I'll look at the ingredients of a RCT that make it such a well-regarded method, and introduce the basic statistics that are used to analyse the results.

A RCT is effective simply because it is designed to reduce many of the systematic biases that were covered in previous chapters. The inclusion of a control group ensures that we can distinguish genuine differences in outcome linked to the intervention from other reasons for change over time (Chapter x). Randomisation of participants to intervention and control groups avoids bias caused either by participants' self-selection of intervention group, or experimenters' determining who gets what treatment. In addition, as noted in Chapter x, where feasible, both participants and experimenters may be kept unaware of who is in which treatment group, giving a double-blind RCT.

RCTs have become such a bedrock of medical research that standards for reporting them have been developed. In Chapter x we saw the CONSORT flowchart, which is a useful way of documenting the flow of participant through a trial. CONSORT stands for Consolidated Standards of Reporting Trials, which are endorsed by many medical journals. Indeed, if you plan to publish an intervention study in one of those journals, you are likely to be required to show you have followed the guidelines. The relevant information is available on the web and can be readily found using a search engine.

For someone starting out planning a trial, it is worth reading the CONSORT Explanation and Elaboration document (Moher et al, 2010), which gives the rationale behind different aspects of the CONSORT guidelines. This may seem rather daunting to beginners, as it mentions more complex trial designs as well as a standard RCT comparing intervention and control groups and assumes a degree of statistical expertise (see below). It is nevertheless worth studying, as adherence to CONSORT guidelines is seen as a marker of study quality, and it is much easier to conform to their recommendations if a study is planned with the guidelines in mind, rather than if the guidelines are only consulted after the study is done.

10.1 Statistical analysis of a RCT

Many of those who train as SLTs or other allied health professions get little or no statistical training. This is unfortunate, as a grasp of basic statistics is essential to understand the research literature on interventions. In some cases, SLTs may have access to expert advice from statisticians, but I suspect that is rarely the case. We have, therefore, a serious dilemma: those who have to administer interventions do not have the skills to evaluate their effectiveness.

I do not propose to turn readers of this book into expert statisticians, but I hope to instil a basic understanding of some key principles that will make it easier to read and interpret the research literature.

The basic analysis of a RCT doesn't need to be complicated. For a simple comparison of intervention vs control groups, we have two sets of numbers on an outcome measure: scores from the control group and scores from the intervention group. Figure 10.1 shows three different ways of reporting results from two groups from a fictitious study that compares the impact of vocabulary training on children's word knowledge in a sample of 4-year-olds. Allocation to intervention was randomised, and thirty children received individualised intervention over a period of three months whereas the remainder received an equivalent amount of time with the therapist having 'business as usual', but without a focus on vocabulary. The primary outcome measure was raw score on the British Picture Vocabulary Scale. This had also been administered at baseline, but we will focus first just on the outcome results.

Table 10.1

Table 10.1 shows the mean and standard deviation for each group. The mean is the average, obtained by just summing all scores and dividing by the number of cases. The standard deviation gives an indication of the spread of scores around the mean. It is a key statistic for measuring an intervention effect. Figure 10.1 illustrates why this is the case. Here we see results from three studies. In each of them the group means are the same, but the standard deviations are very different. Study X shows an impressive group effect with little overlap between the groups; study Z shows a distinctly unimpressive group effect where there is substantial overlap, and study Y – which corresponds to Table 10.1 is intermediate. Very often with research results, we have something resembling the scenario shown in study Y: one mean is higher than the other, but there is overlap between the groups. Statistical analysis gives us a way of quantifying how much confidence we can place in the group difference: in particular, how likely is it that there is no real impact of intervention and the observed results just reflect the play of chance.

Figure 10.1

Understanding the notion of chance is key to understanding statistics. Suppose we have a pack of cards and I deal each of us a hand of 10 cards, and we total the points on our cards. Assuming we have a fair pack of cards, we can know what the pattern of results would look like if we re-ran this experiment 100 times. It is shown in Figure 10.2.

In this chapter, I'll be showing you how you can simulate data like this using R. I find it very helpful to do this to gain understanding of statistics. But you really don't need to do this if you prefer not to. You can just ignore the bits of text that show R scripts and focus on the results.

The reason for doing this simulation is that it emphasises that, even when there is no difference between us – I have a fair pack of cards and I'm not cheating, sometimes I'll get a much higher score than you. And if we keep running the simulation for say, 10,000 times, we can work out just how many times I get a score that is at least 10, 20 or 30 points higher than yours: this occurs on x%, x% and x% of occasions.

Suppose that I now introduce you to my friend, Zac, and you find that he on 20 rounds he wins by a 10 point margin every time. Is he just lucky, or is he cheating? Well, you can work out the odds using probability theory – i.e. with cards we know exactly how likely different hands are, and so we can work this out. Would you be justified in accusing him of cheating? You can never be certain. It could just be a fluke, but the lower the probability of his card hands, the more persuaded you will be.

Interpreting data from a trial is just like this. There are various approaches taken by statisticians, who can argue vehemently for the superiority of one method over another. I'm only going to cover the approach called Null Hypothesis Statistics Testing (NHST) because it is widely used and straightforward – even though many statisticians regard it as inferior to alternative approaches. But for determining whether an intervention is or is not effective, it is a reasonable way to start out.

Statisticians often complain that researchers will come along with a collection of data and ask for advice as to how to analyse it. A point that is sometimes missed is that you can't know what statistics to use unless you have a hypothesis. In the case of an intervention trial, the hypothesis will usually be 'did the intervention make a difference?' There is, in this case, a very clear null hypothesis – that the intervention was ineffective, and the outcome of the intervention group would have been just the same if it had not been done. The null hypothesis

testing approach answers just that question: it tells you how likely your data are if the the null hypothesis was true. To do that, you compare the distribution of outcome scores in the intervention group and the control group to see if the difference in the average of the two groups is any greater than you'd expect given the variation within the two groups. This is exactly what the term 'analysis of variance' refers to.

An intuitive way to look at this is to consider that you can compute three means and standard deviations for each study shown in Figure 10.1.

and work out how much variability there is between the two groups relative to the variability within each group.

There are

10.2 Obfuscation and omission in the reporting of results

Chapter 11

Yet more bias: distortions arising at the analysis stage

Dangers of too many outcome measures Dangers of post-hoc subgroups and covariates

Chapter 12

False positives and false negatives: are we missing true effects?

The aim of many research studies is to investigate whether there is evidence for a specific hypothesis. We typically want to test this hypothesis in a robust way that gives us a clear indication of evidence for or against. Study design is paramount in ensuring that we can trust our results and have confidence in any inferences drawn. The implications of poor study design are not easily visible as many statistical tests will generate results regardless of whether certain test assumptions are met, or if we are using a small sample. This failure to meet certain assumptions or using an insufficient sample size can lead to an increase in false positives, also known as a Type I errors (α), or an increase in false negatives, also known as Type II errors (β). Some more formal definitions with examples will help us to clarify these concepts.

12.0.1 Type I error (α)

A false positive occurs when the Null hypothesis is rejected and a true effect is not actually present. Typically, the type I error rate is controlled by setting the "significance criterion" at 5% or sometimes more conservatively 1%. In real terms, this would indicate that if $\alpha = 0.05$ and a true effect is not present, we would expect 1 in 20 results to be statistically significant but in fact this will have occurred by chance alone.

For example, Researcher X wants to test the hypothesis that eating carrots improves children's IQ by 2 points. The researcher assigns 6 children to two groups, one group eat carrots and the other group eat their usual snack. The data is collected and a significant finding is found (Null hypothesis that carrots do not cause a 2 point increase in IQ is rejected). The researcher is jubilant that he has discovered an interesting result for publication.

His sceptical colleague decides to check this result and decides to try to replicate the finding. The colleague re-runs the study with a new sample of children, but does not alter any other details of the study. Unfortunately, the colleague does not replicate the finding. The first result is likely to be a false positive for a number of reasons including sampling variability in small samples, small effect size, or even just poor quality data.

12.0.2 Type II error (β)

A false negative occurs when the Null hypothesis is accepted and a true effect is actually present.

Now consider a second example, where the sceptical colleague wants to test a well-established and frequently-replicated result. she wants to confirm whether on average children who read more frequently obtain better exam results when they are tested on vocabulary. Once again, 6 different children are split into two groups, one group reads with a parent each night for a month and the other group continues with their usual routine. The study is

Table 12.1: Confusion Matrix for binary test

		True	
		Condition positive	Condition negative
Predicted	Predicted Positive	True Positive POWER	False Positive Type I error
Predicted	Predicted Negative	False Negative Type II error	True Negative

run and the Null hypothesis is accepted as the statistical test results in a non-significant p-value (P>=0.05, in this case).

The sceptical colleague is confused as she cannot replicate a known effect, but there is overwhelming evidence to indicate that an effect should be present. We look to the study design and oncemore, we see that the sample size was small and sampling variability will have played some role in causing a type II error.

Table 12.1 presents a contingency table of the potential combinations of test or predicted condition and true condition. This is a common representation of the classifications of power, false positive, false negative, and true negative for a binary test classifier.

12.1 Statistical power

As we have seen in our earlier examples, a frequent concern in study design refers to sample size and it's related concept of statistical power. Power can be defined as the probability that some statistical test will reject the Null hypothesis given that an alternative hypothesis of interest is true.

Consider the following example, we want to investigate whether there is a significant difference in mean scores between two groups, perhaps an intervention and a control group. The indivduals are randomly assigned to each group and for simplicity we assume that there are no additional differences between the groups. The intervention is administered and we collect data on all individuals before running our statistical analysis. The power is then a function of several quantities:

- Sample size
- "Significance criterion", also known as the Type 1 error rate (α)
- Size of the effect of interest
- Type II error (β, note that Power=1-β)

We can show the relation between these quantities of interest visually through an example using the z-test which is a simple one sample location test. The idea of this test is to assess whether a sample of some quantity of interest has the same sample mean as the population of the same quantity. In the figures below we will use a one-sided z test, which indicates that we have a direction specific hypothesis and is one of the simplest tests to introduce these concepts. A side note, two sided tests are more common in practice due to the fact that it is more realistic to test for a difference regardless of direction of effect.

The first figure 12.1 shows the case for a single individual, n=1. We have set the standard deviation of both Null and Alternative distributions to equal one, and their means to zero and one respectively. When we have a sample size of 1 individual, we see that a small area is defined for power, indicating lower power and higher Type II error.

The second figure 12.2 presents the same one-sided z test but here the sample size has increased. We should notice that two things have appeared to change. Firstly, we see a greater distinction between the two distributions. Secondly, we see that the critical z value (vertical dashed line) has changed location. The distributions have not changed their location (peak of each bell shaped curve), but the spread of each distribution has shrunk as a result of the sample size as the spread is directly proportional to the sample size. The reason for this change lies in the formula for the standard error and Z score test statistic itself,

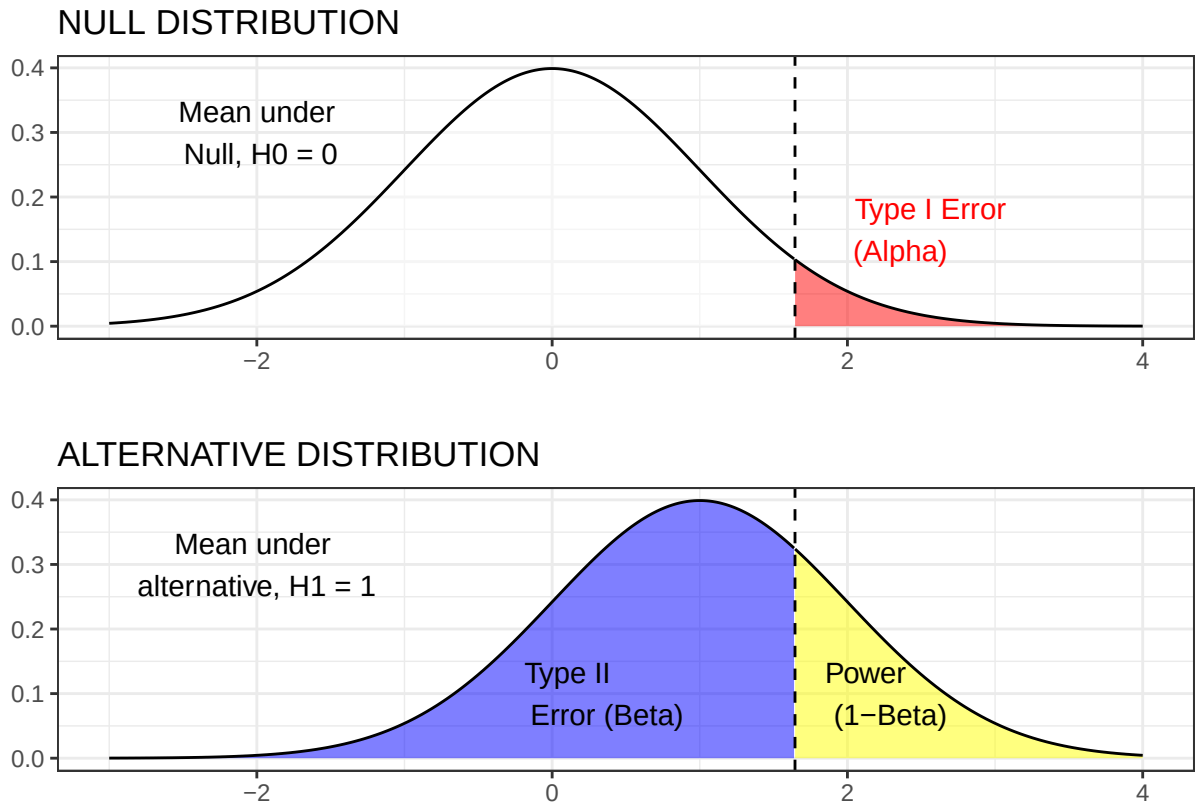


Figure 12.1: Z test: statistical power, N=1

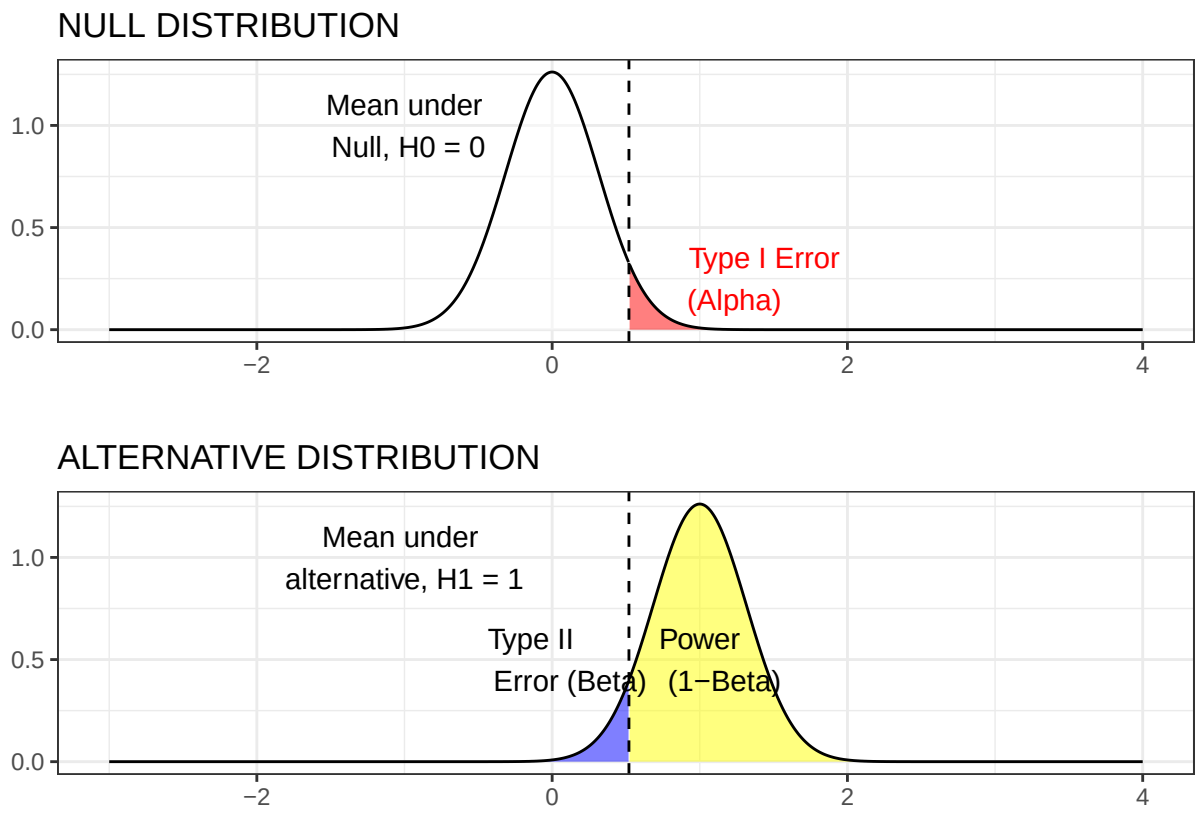


Figure 12.2: Z test: statistical power, N=10

$SE = \frac{SD}{\sqrt{(n)}}$, where n is the sample size, and SD is the standard deviation.

The z value or score is also proportional to the sample size, so we see that this value changes accordling. The z score is defined as follows,

$z = \frac{M-\mu}{SE}$, where M is the sample mean score, μ is the population mean, and SE is as defined above.

The shaded areas on the density plots directly relates to the concepts outlined above: power, type I, and type II errors. When the sample size increases, the standard error (SE) and z score both reduce. The only input parameter that has changed between the two figures is that the power has increased. We notice that the type I error rate (area in red) is proportionally the same at 5% and the effect size remains fixed, so we see a change in the only two remaining quantities, power and type II rate. This is because these quantities are linked. The area under the density curve must always remain at 1, so proportionally, we can calculate the power as 1-(type II rate, β) and vice versa. We can visually see this in both figures by looking at the specified areas for the alternative distribution.

An important point to note is that each of the following elements: power, sample size, effect size, and type I and II errors are intertwinned with each other. Hence, if one element is changed this has some effect on the other quantities. It also gives an insight on how we can use this knowledge to design our studies to minimise type I and II errors and increase statistical power.

Typically, clinical trials in medicine will design studies to achieve 80% statistical power and depending on the statistical analysis strategy, will employ some method to control type I error rate or use a more conservative rate (traditionally in NHST $\alpha = 0.05$). Fixing these two quantities leaves us with a smaller number of elements to

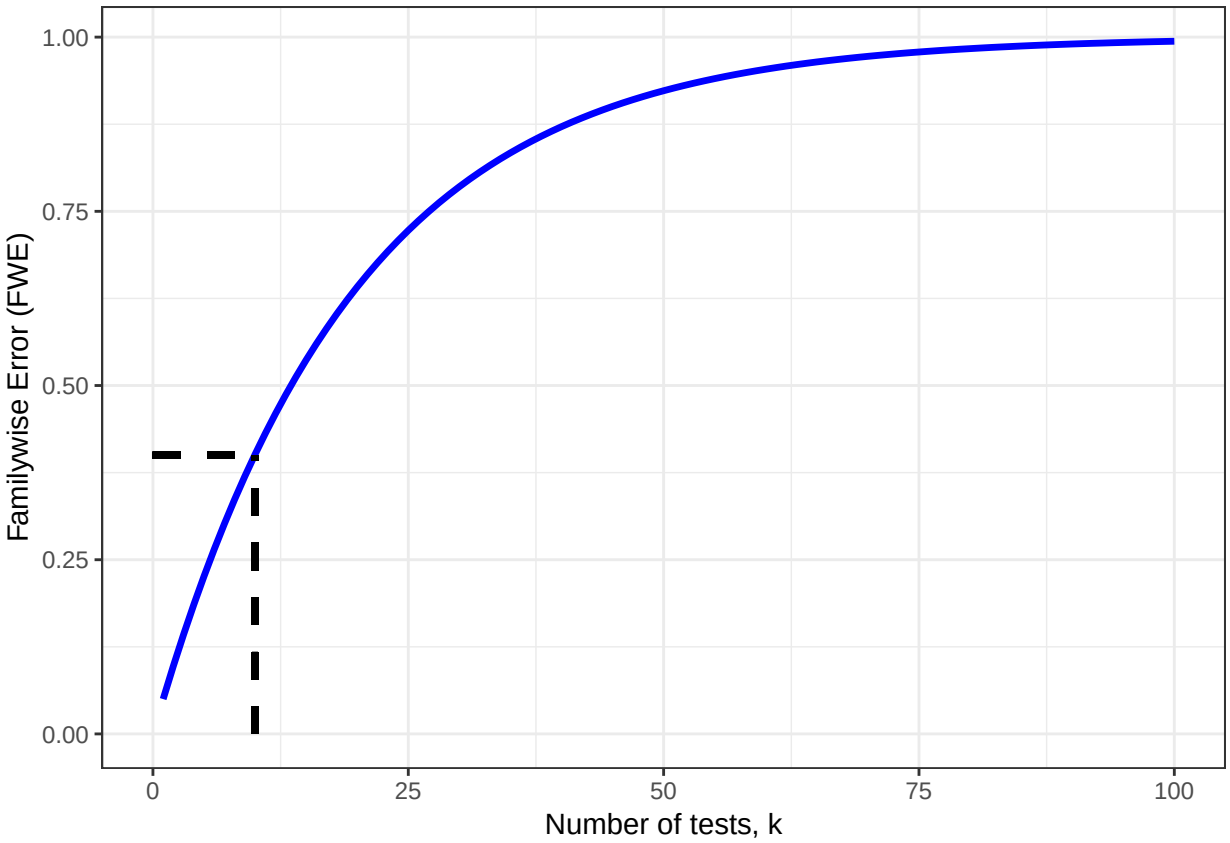


Figure 12.3: Plot of relationship between familywse error rate and number of statistical tests

vary, but ususally we are designing to answer a particular scientific question which will also provide theoretical constraints. More specifically, an effect size of interest will be dictated ideally from a meta-analysed estimate of the effect size, or designed to the smallest effect size of interest that is practically or clinically meaningful. This leaves us with sample size as our only remaining element.

A power calculation will vary sample sizes for a specific statistical analysiswhilst keeping statistical power, type I error rate, and effect size of interest at constant values. Sample size is usually most influenced by the size of effect and researcher constraints, i.e. funding or time. A large effect size will require fewer unique observations and conversely, smaller effect sizes require very large samples. The logic behind this can be seen in figure 12.2, if we increase the effect size, the central tendencies (mean) of each distribution diverge, so the overlap of the distributions becomes less, provided that the standard errors remains unchanged.

12.2 Multiple hypothesis testing

Even when we have a well-designed and adequately powered study, if we collect multiple outcome variables or if we are applying multiple tests to the same data, then the outcomes are unlikely to be completely independent, so we increase our change of finding a false positive. Returning to the idea that we have a type I error rate for a single test at 5%, then suppose we apply k tests to our data. We increase our false positive rate dramatically for the set of tests, officially known as the family-wise error rate (FWER). The new significance level is given by $1-(1-\alpha)^k$ (i.e. the probability of finding at least one false positive). If we assumed that we have performed 10 tests (k=10), we have increased our chance of obtaining a false positive to approximately 40%.

Figure 12.3 shows the relationship between the familywise error rate and the number of tests administered to the data. Clearly, the more tests applied leads to significant increase in the chance of at least one false positive result. There are many different ways to adjust for the multiple testing in practice. We shall discuss some of the most commonly applied.

12.2.1 Bonferroni Correction

The Bonferroni correction is both the simplest and most popular adjustment for multiple testing to protect the type I error rate. The correction works by dividing the type I error by the number of tests conducted.

For example, say we had some data and wanted to run multiple t-tests between two groups on 10 outcomes which all measure a similar effect of interest. The Bonferroni correction would adjusted the α level to be $0.05/10 = 0.005$, which would indicate that the critical α would be $1-(1-\alpha_{adjusted})^n = 1-(1-0.005)^{10} = 0.04888987$ which is approximately close to our original α as required, so we have successfully controlled our type I errors at approximately 5%.

Bonferroni is widely used due to its simplicity but it has the consequence that it can be overly conservative. We say that a test is overly conservative if we over correct in some cases which increases our type II errors and reduces our statistical power. This is often the case when we have some dependency between the outcomes

12.2.2 False Discovery rate (FDR)

The false discovery rate differs from the Bonferroni as the correction controls the expected proportion of false positives rather than all false positives.

$FDR = \frac{False\ positive}{False\ positive\ +\ True\ positive}$

We will present two common techniques for FDR procedure, Benjamini-Hochberg (BH), and Benjamini-Yekutieli (BY). The BH procedure can be summarized into four steps:

1. Sort a set of p-values from a set of multiple comparisons into ascending order.
2. For each p-value, assign a rank. For example, the first p-value could be 1 and the second would be 2, and so on.
3. Next we need to calculate a BH critical value for each p-value. We use the formula, $(i/n)\alpha$, where i is a particular p-value's rank; and n is the number of tests.
4. Original p-values are now compared against the BH critical values (3), finding the largest p-value that is smaller than the critical value.

We can solidify this procedure by looking at an example. Consider again that we revisit the set of t-tests from our Bonferroni correction. We are sure that the correction is overly conservative as some of the test are moderately correlated with each other.

Table 12.2 shows the p values obtained from 10 t-tests on data obtained from the same subjects but different outcome measures. If we considered the p values are sufficiently indpendent, we could conclude that the first six variables(x1-x6) are statistically significant, but on further inspection and deciding to correct using the BH procedure, we find that the we find only X1-X4 are statistically significant under a more robust FWER (green coloured cells). Had we chosen the more conservative Bonferroni correction, we would only find X1 and X2 remain statistically siginficant ($\alpha_{Bonf} = 0.05/10 = 0.005$, blue coloured cells).

The second method that controls the False discovery rate is the Benjamini-Yekutieli (BY), which is a more conservative correction but allows for dependence between the tests (i.e. the tests are correlated; therefore, the p-values will also be correlated). The adjustment resemmbles the BH procedure with an additional constraint,

$(i/n)(\frac{\alpha}{\sum_{i=1}^{k} 1/i})$

Table 12.2: Corrected p-values using several different methods (Bonferroni, BH, BY)

Variable	Rank	alpha	p.value	BH	BY
X1	1	0.05	0.002	0.005	0.005
X2	2	0.05	0.004	0.01	0.007
X3	3	0.05	0.008	0.015	0.008
X4	4	0.05	0.012	0.02	0.01
X5	5	0.05	0.023	0.025	0.011
X6	6	0.05	0.041	0.03	0.012
X7	7	0.05	0.054	0.035	0.013
X8	8	0.05	0.091	0.04	0.015
X9	9	0.05	0.12	0.045	0.016
X10	10	0.05	0.2	0.05	0.017

Table @ref(tab:BY_example) shows the p values that remain statistically significant under the Benjamini-Yekutieli adjustment (coloured in red). In our example, the BY is more conservative than the BH but is less conservative than the Bonferroni.

12.2.3 Permutation methods

Resampling techniques offer a new perspective on correcting for multiple testing as the basis is not according to adjusted p-values. If we recall in 12.1, we have a null distribution which is crucial for obtaining our test statistic. The permutations approach creates an empirical estimate of this null distribution by re-sampling, say M times, the total number of observations, in a population sample. The benefit of this approach is that the statistical significance is estimated from the data. This also has the added benefit that properties of the data including irregularities are carried through in the permuted data sets. The general procedure is as follows:

1. Using the original data set, a test statistic and corresponding p-value are calculated.
2. Next, the data are permuted (resampled), so we obtain a jumbled sample of the original data set but with the allowance of resampling with replacement. Using this new permutated sample data, we calculate the test statistic and associated p-value.
3. Step (2.) is then repeated M times (typically, $>$1000 repetitions), and each time the test statistic and p-value are recorded to form empirical distribution of the test statistic.
4. The permuted distribution is then calculated as the frequency of M permutations that the original data test statistic is smaller than each permuted test statistic, then dividing this frequency by M, the number of permuted iterations.

The same statistical test is used according to the original design, but this test is fitted multiple times. The permutations randomly sample values from each group, as the observations from each group are considered as exchangeable. The assumption in the independent groups test is that the level of significance is not dependent on particular pairings of observations from each group.

Suppose we have two groups of individuals and have measured language performance using three different instruments. We choose to use the Student's two independent samples t-test sequentially to test for differences between the two groups of individuals on each of the three measures. We have chosen not to use a correction that adjusts the p-values, instead we decide that a permutation approach will be more robust. Figure @ref(fig:permutation_test) shows the result of our analysis using the permutation corrected t-tests. The cross points are the actual p-values of the t-tests on the original data. The 95% confidence intervals are constructed from the permuted distributions.

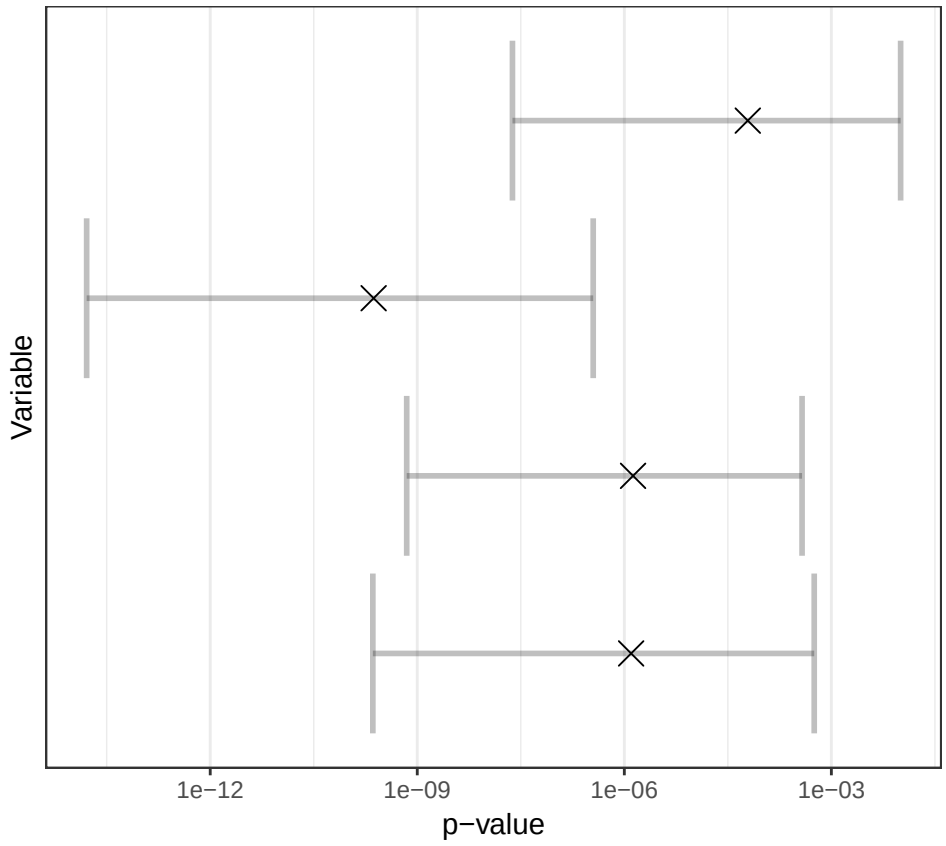


Figure 12.4: (#fig:permutations_test)Permutation distributions for individual tests

Chapter 13

Drawbacks of the RCT

Transfer to real life: efficacy and effectiveness Inefficiency: need for unfeasibly large samples Heterogeneity and personalised intervention

Chapter 14

Alternatives to RCT: regression discontinuity

The regression discontinuity follows a quasi-experimental design and is typically used to avoid confounding bias (i.e. when another variable is present that distorts the effect of intervention) when assesssing interventions. It is particularly useful when assignment to intervention is based on some pre-specified threshold from a continuous measure. For example, let us assume that we wish to assess a new reading intervention in a school. Students are assigned to the intervention group according to a general language test. Students that score below a certain threshold are assigned to the intervention group, and rather than discrd their peers that scored above the threshold, they forma natural control group. The experimenter is not therefore required to randomly allocate students with poor language scores to the control group when they might benefit from the potential effect of the intervention if it is found to be effective.

Before we delve into the regression discontinuity designs, it may be useful for us to first take a look at general linear models, in particular, ordinary linear regression. Linear models consist of some of the most commonly used methods in applied research including ANOVA, ANCOVA, linear regression, t-tests, F-tests and MANCOVA.

Linear regression's purpose is primarily to understand the straight line relationship that exists for some bivariate data. We assume a causal direction inthe relationship unlike correlation as we specify an outcome (dependent) variable and a predictor (independent) variable. The idea being that as we change the value of predictor we see the effect that this has on the outcome variable, by quantifying the relationship according to some statistical model. Figure 14.1 (A, left) shows some bivariate data which appears to have an underlying linear relationship

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 (X_{i,1} >= c)$$

where

Statistical issues encountered RDD analysis:

1. Incorrect specification of regression form (regression is non-linear or needs a polynomial form)
2. Misallocation of treatment
3. Inadequate statistical power (sample size too small)
4. Limited generalization of effects

14.1 Mediators

14.2 Moderators

If we return to our linear regression definition from earlier, we have some predictor variable X that has a linear causal relationship with Y, i.e. X causes Y. Moderation occurs when a third variable is introduced, say M that

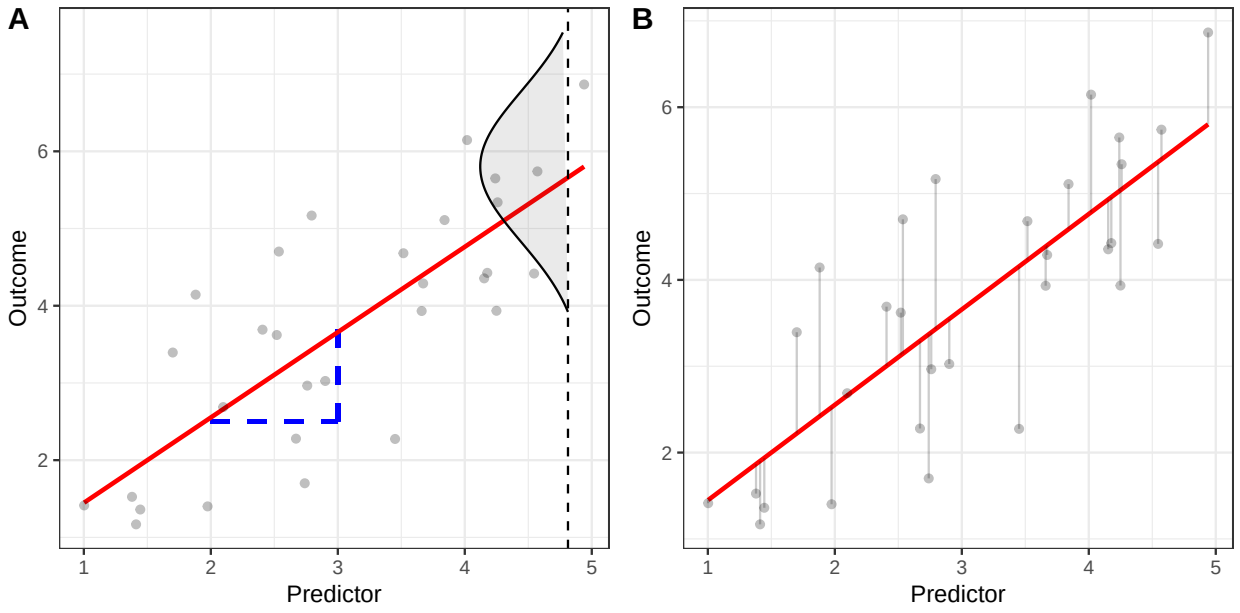


Figure 14.1: Linear regression: Plotting residuals, intercept, and slope explainers

changes the strength of causal relationship between X and Y, hence moderating the relationship.

Figure 14.4 shows a disgrammatic representation of the moderator effect. The moderator is added into the statistical framework as an interaction. The interaction is added into the regression model formula as an extra variable with a new associated parameter estimate. The extra variable is simple the two interacting terms multiplied together. We can give an exmaple of this by returning to the regression equation that we met earlier. The interaction appears as the multiple of X_1 and X_2.

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 (X_{i,1} X_{i,2})$$

We can visual the effect of the interaction much more clearly by looking at some simulated data. Plotting the data, for two types of combinations of variables. Figure 14.5A (LEFT) shows the relationship between a nominal Moderator and a continuous predictor. Figure 14.5B (RIGHT) shows the relationship between a nominal Moderator and a nominal predictor.

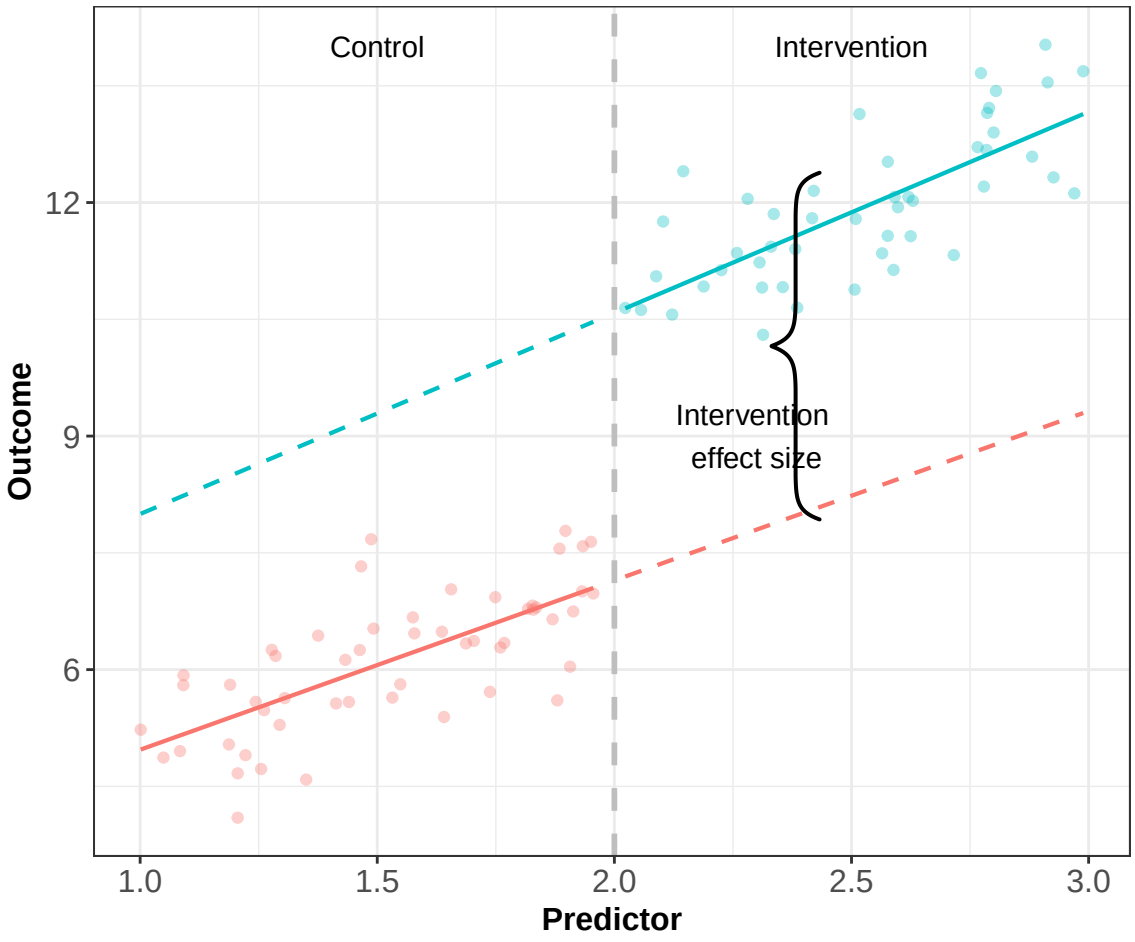


Figure 14.2: Linear regression: Plotting residuals, intercept, and slope explainers

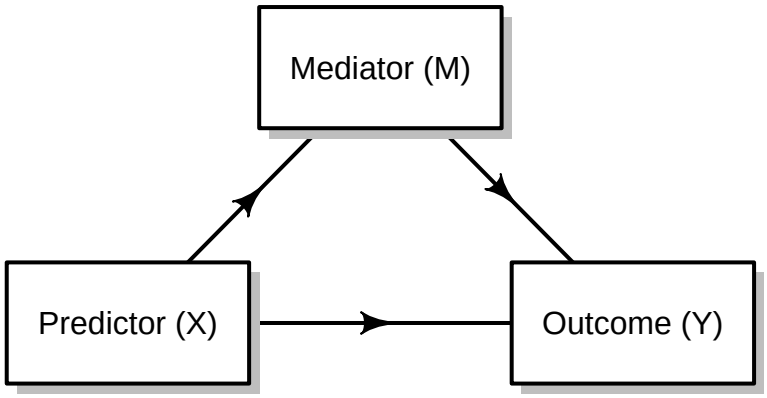


Figure 14.3: Mediator path diagram

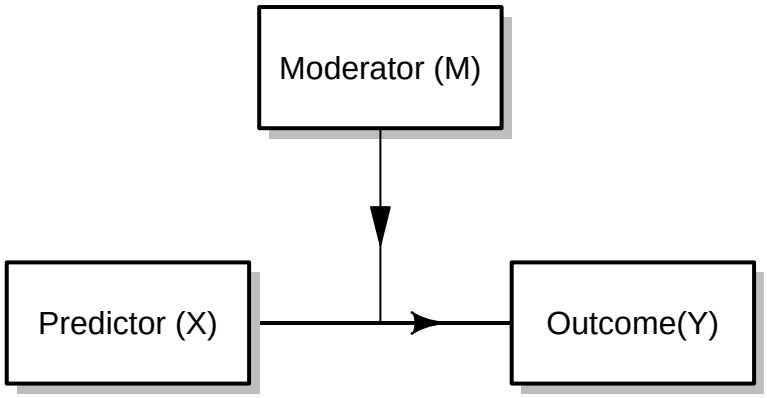


Figure 14.4: Moderator path diagram

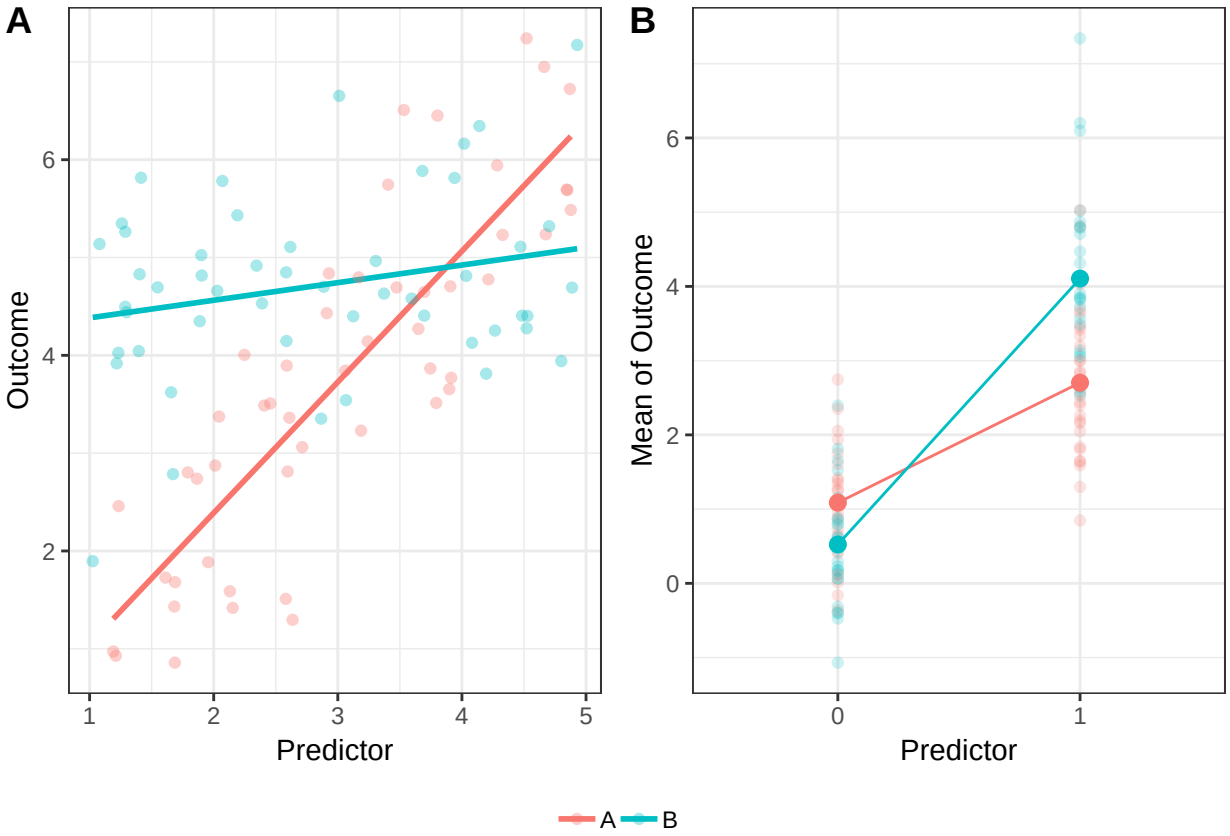


Figure 14.5: Plotting the interaction effect in regression. (A) shows the relation between a nominal Moderator and continuous predictor; (B) shows the relationship between a nominal Moderator and nominal predictor.

Chapter 15

Alternatives to RCT: within-subject designs

15.1 Single case designs

When a single case design won't work

Chapter 16

Adaptive Interventions

16.1 Just-in-Time adpative interventions (JITAI)

16.2 Micro-Randomized Trials (MRT)

16.3 Sequential, Multiple Assignment, Randomized Trial (SMART)

Chapter 17

Practical obstacles to the ideal study

17.1 Sample size: need for team science?

17.2 Over-regulation of research: when ethics committees misfire

17.3 Problems in generalising to the real world

Chapter 18

Can we believe the literature? Publication bias

Chapter 19

Pre-registration as a means to combat publication bias

Chapter 20

Avoiding waste: the need to start with a systematic review

Chapter 21

A template for a research protocol