Randomised trials in education: An introductory handbook

Professor Carole J Torgerson

School of Education

Durham University

carole.torgerson@dur.ac.uk

Professor David J Torgerson

Director, York Trials Unit

Department of Health Sciences

University of York

david.torgerson@york.ac.uk

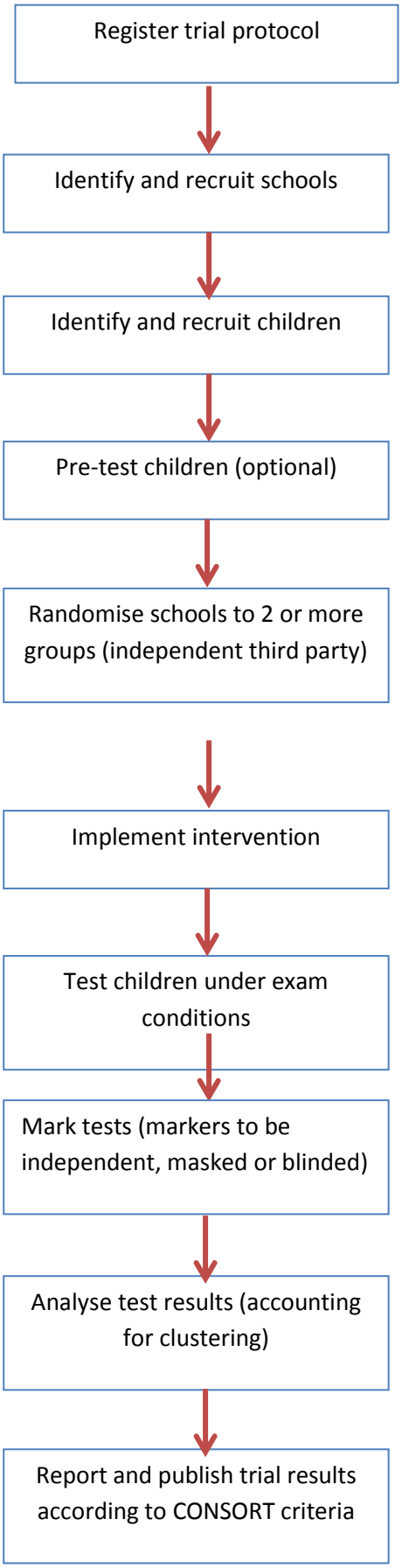
NB: This is work in progress. Any comments on this draft are welcomed. Please email them to one of the authors above.

Introduction

Randomised controlled trials (RCTs) are the best approach for demonstrating the effectiveness of a novel educational intervention. Most other approaches are susceptible to selection bias. Selection bias occurs because the method of selecting schools or students to receive an intervention is related to outcome. Consequently, any difference we observe between a group receiving an intervention and a group not receiving the intervention may be due to the selection characteristic rather than the intervention itself. A RCT escapes the problem of selection bias at the point of group formation by using the process of random allocation to produce the allocated groups. Because each school or student has the same probability of getting the intervention or not this abolishes selection bias. However, whilst random allocation is a necessary first step towards producing a robust RCT, it is not the only process that needs to be done rigorously to ensure an unbiased result. In this introductory hand-book we outline the main issues an evaluator needs to consider when designing and conducting a rigorous RCT.

Table of contents

Key steps for a successful school-based RCT



Trial registration

All trials should be registered before randomisation. This can be done at www. ISRCTN.org

Description

Registration of trials means putting them in a database that is publicly accessible and can be searched *before* the trial is completed.

How?

Register with an online publicly available database. Several have been established, mainly for health care trials, and at least one of them – the International Standard Randomised Controlled Trial Number Register - will also register non-health care trials. If you go to the www.ISRCTN.org website you will find detailed instructions on how to register your trial for a modest payment.

When?

Registration should be undertaken as soon as possible once it has been agreed that the trial will be undertaken. Ideally this should occur *before* any schools or pupils are recruited into the study.

Why?

There is good evidence from methodological research that those trials that have not been published are those that find either no difference between intervention or control conditions or, a difference that favours the control group. For example, in a systematic review of trials of phonics instruction for the teaching of reading there was evidence that small trials showing no difference were not published (Torgerson, 2003). Registration of trials should help prevent selective publication.

Concealed randomisation

Random allocation must be undertaken by someone who is, and is seen to be, independent from the development and delivery of the intervention

Concealed randomisation is where all participants, including pupils, schools, those undertaking the recruitment and those who developed or who will implement the intervention do not know which group the schools, classes or pupils are randomised into until after this has happened. Therefore, there is no foreknowledge of the randomised allocation.

How?

The key issue with respect to randomisation is that it must be *independent*. The randomisation sequence (the sequence of allocations, such as ABAABABABB etc.) needs to be generated by a robust process and the allocation to the groups A or B must be concealed from the person recruiting the schools or pupils until after the allocation has been done. Independent and concealed allocation can *always* be achieved. Do not confuse independent, concealed allocation with blinding (which is more difficult to achieve). Independent allocation can be done by recruiting, for example, an independent statistician or data manager who can randomise the schools, classes or lists of pupils. Some trials use internet, telephone or other distant allocation methods: again these are provided by an independent organisation. For example, most clinical trials units in the UK provide randomisation services, which typically take the form of a web-based system, where the details of the school or pupil are entered onto a database, which then provides the allocation. Importantly, the software 'locks' the allocation and prevents any post-randomisation changes. Some form of computer generated randomisation should be used. Some clinical trials units provide this service for education and social science trials.

When?

Randomisation of schools, classes or pupils must be done *after* pupils/parents/teachers have been recruited and given their consent to participate in the study.

Why?

The first big problem when we come to randomisation is the potential for allocation mishaps. Consider a coin tossing approach. We may toss a coin to allocate a pupil or school to an intervention and find that the coin toss results in that particular pupil or school being allocated to the control group. If a school had agreed to enter a trial in the hope of getting the 'intervention' and found that that randomisation had allocated them to the control group then there is the potential for the researcher to re-toss the coin in order to get the 'right' allocation for the school. If this is done it will lead to bias. Furthermore, the researchers themselves may attempt to subvert the randomisation for their own reasons. If such manipulation of the allocation occurs

then the trial is no longer 'randomised'. Misallocation or subversion of the allocation is not a 'theoretical' problem. Of all the threats to the internal validity of RCTs substandard randomisation is probably the one with the most evidence of threat to rigour in design. Most of the methodological evidence, however, originates from health care trials (e.g., Schulz et al, 1995; Hewitt et al, 2005), although there is some evidence from non-health care trials, which suggests the problem is not confined to health care researchers (e.g., Boruch 1997). There have been numerous methodological papers over the years and the consensus of these papers is that, on average, when a trial uses an allocation method that can be tampered with by someone with a vested interest in the outcome of the trial then the findings of the trial are more likely to support a positive result compared with trials where the allocation has been undertake independently.

A note is included here about what is a 'vested' or conflict of interest. Whilst the developer of an intervention may have a financial interest in the outcome of a given trial, of equal importance is the potential for the researchers to have an intellectual interest in the outcome of the trial. Or simply there may be an overwhelming feeling by the researcher who developed the intervention that it must work; the need to prove this to others, who are sceptical, might lead to the undermining of the allocation.

Methods of randomisation

Simple randomisation with small sample sizes (i.e., <100 schools or students allocated) might lead to unacceptable imbalances between groups and a restricted randomisation with stratification is needed. For class randomised trials teachers should be linked to the class before randomisation.

Description

There are broadly two categories of randomisation: simple or restricted allocation. Simple randomisation is akin to tossing a coin, whereas restricted randomisation puts limits on the probabilities to ensure that certain characteristics (e.g., school size) are evenly distributed across groups. Stratification means that the randomisation process is governed by some characteristics of the sample such that we can ensure balance on that variable(s), which simple randomisation does not guarantee.

How?

Approach an independent person to undertake the randomisation. Many computer programmes can do the allocation.

Simple randomisation

A list of participants and the computer programme divides them into (usually) equally sized groups at random. Or if there is recruitment over time the participants are individually randomised as they present themselves to the researchers, by a computer programme.

Blocked stratified randomisation

If the sample size is small (n <100), a better approach at producing balanced groups is through blocked, stratified randomisation. If we wanted to stratify on size and we had 20 schools we would divide the 20 schools into two groups on the median school size (the stratification variable), with 10 schools above the median and 10 below. We could then, within each block of 10 schools, randomly allocate 5 large schools to the intervention and 5 to the control; similarly, we could randomise within the 10 small schools - 5 to the intervention and 5 to the control condition. If we had 21 schools one of the groups would have 11 schools and we would ensure that we produced a block of 11 allocations where five were in one group and six were in the other (the larger or smaller group determined by chance). However, if we start to stratify on several variables there could be a problem - too few schools, or students, with the stratified characteristics to ensure numerical balance across the different factors. Therefore, it is wise to only stratify on one or two variables only or use minimisation.

Minimisation

One approach that allows better balance than stratified blocked randomisation with small sample sizes is a deterministic method known as minimisation (Torgerson and Torgerson, 2008). Given a list of 20 schools the first school is allocated at random. Subsequent allocations are made by using an algorithm which places the next school into the group where the differences of existing allocated schools, or pupils, are minimised between them. This process can deal with a larger number of variables than blocked, stratified randomisation.

Matched pairs randomisation

This is when schools (or pupils) are matched on some characteristic (e.g., school size) and then one is randomised to the intervention and one is randomised to the control condiiton. In some circumstances matched pairs randomisation can improve the power of the study. However, there are drawbacks (see below), which usually mean other approaches are better.

Pairwise randomisation

This is when pairs of schools, unmatched by characteristics, are randomised. This ensures numerical balance between groups, but not necessarily balance in school characteristics. It may be advantageous to do this for logistical reasons, if for example, the intervention implementation can only cope with one school at a time and we wish to avoid too many schools getting it too quickly. This method has some advantages over matched pairs randomisation.

Linked Class randomisation

If the unit of randomisation is class and the intervention is delivered by the class teacher then the teacher needs to be linked to the randomised class before allocation occurs.

Why?

Many educational trials are small when classes or schools are the unit of randomisation. For instance if we were to randomise 20 schools by using simple randomisation (i.e., coin toss) we may easily end up with 15 schools in one group and only 5 in the other group or even greater numerical imbalances. More worryingly, even if we ended up with exactly 10 in each group we might, by chance, have the 10 largest schools in one group and the 10 smallest in the other. If educational outcome were affected by school size, which it might well be, we could not then differentiate between any intervention effects caused by the intervention and those caused by school size. Therefore, for small sample sizes alternative approaches are required.

Matched or paired randomisation is often used in educational trials. For example, if you have 20 schools you might rank them in order of size and take the first two schools and randomly allocate one to the intervention and the other to the control condition; you would continue this process until all of the schools have been allocated. This will result in the groups being matched on size of school. There are a number of problems with this approach, however. The statistical analysis needs to take the matching into account and this can lead to loss of power and some loss of flexibility in the analysis. Also if one of the schools dropped out after randomisation this could lead to problems in the analysis. Furthermore, if you had 21 schools, what

would you do with the unpaired school? Some may choose *not* to include the 21st school, which would be a waste as we would have lost an observation that would have increased the study power and perhaps the generalizability of the experiment. Generally, therefore, it is best to use alternative approaches to achieving balance in the randomisation, such as stratified randomisation or minimisation.

In terms of linking classes to teachers before randomisation this is important. There are two sources of potential selection bias in a class-based randomised trial. There are underlying differences between classes in terms of the characteristics of the children. This is dealt with by randomising the classes. The other source is teacher differences. To ensure that this is also balanced between groups teachers must be 'linked' to the class before randomisation, then the randomisation process ensures balance in terms of class composition and of teacher characteristics. If the teachers are not linked before randomisation, then, although the randomisation ensures balance in terms of child characteristics, teacher level selection bias could occur if teachers select themselves to teach the intervention classes after randomisation. Thus, the strongest (or weakest) teachers may volunteer to teach the intervention in preference to teaching the control condition: if this happens the trial's results will be biased.

Pre-testing

A pre-test is usually a good idea if it is possible. Aggregated pre-test data are nearly as good as individual level data. Pre-testing must be done before randomisation and does not have to be the same type of assessment as the post-test.

Description

Most educational trials ask students or pupils to take a pre-test. This is not a pre-requisite of a randomised trial: indeed, there are many situations where it is not either feasible or desirable to have a pre-test. Nevertheless, if it is possible to do a pre-test then this is usually an excellent idea.

How?

The pre-test does not actually need to be the same test as the post-test. Indeed, it could even be from a different subject area. For example, if we were undertaking a trial with numeracy as an outcome then ideally we would use a similar assessment at both pre and post-test. However, if for cost or logistical reasons it is not possible to set the same assessment we can use other tests, as long as they correlate with the post-test. For example, if a national literacy test had been conducted on the children and these data were available it is likely that this literacy test will correlate quite highly with the post-test maths assessment (although not as well as a national numeracy test). This, then, can be used as a pre-test. Obviously this is an extreme example; it would be better to use a historical maths test which is likely to have a stronger relationship with a post maths test than a literacy test. For example, in a RCT of a maths intervention – *Every Child Counts* – the outcome test was Progress in Mathematics 6, but the pre-test was the Sandwell Maths test, which correlated very strongly with the post-test (Torgerson et al, 2013). The important point here is that for a pre-test to be useful it must have a strong relationship with the post-test. Sometimes we can get pre-test estimates at the school or class level but not at the level of the individual student. Although individual student data are best, when the analysis is at the cluster level then aggregated cluster level data are nearly as good. Thus, data aggregated at the class or school level perform nearly as well as if we had used pre-test data on individual students. This is an important point because often it is possible to get school level data (e.g., national test data is often publicly available by school but not for individual students) at low or no cost, whilst implementing a pre-test at the level of the student may be costly and logistically difficult.

When?

The pre-test must, if at all possible, precede randomisation. This is because if it occurs after randomisation then knowledge of the group allocation may influence the pre-test scores.

Why?

A pre-test adds information to the final statistical analysis. If the pre- and post-test have a strong correlation, say 0.7, which is quite common in educational research, that size of correlation can led to a reduction by about half of the required sample size, or significantly increase the power of the trial.

Compliance and intention to treat

Intention to treat analysis should always be the primary analysis of any RCT. In the presence of non-compliance the best method of adjusting for this is to use Complier Average Causal Effect analysis.

Description

Poor- or non-compliance with the intervention threatens the intervention's effect size estimate by the trial because the estimate is diluted. Consequently, we might see an under-estimate of the true effectiveness and any observed differences may not be statistically significant. There are a number of different ways of dealing with non-compliance in the analysis: some are sensible and others may introduce other biases.

How?

The primary analysis of any trial should be 'Intention to treat' (ITT) analysis. This is where we analyse the schools or students on how they were randomised. If, for example, in a trial of 20 schools one of the schools in the intervention group did not comply with or accept the intervention we would include them in the analysis as if they *had* received the intervention. This analysis gives us a conservative and unbiased estimate of the effect of the intervention being offered to schools. To assess the effect of non-compliance on outcome the most rigorous approach for dealing with non-compliance is to use a statistical technique known as Complier Average Causal Effect (CACE) analysis (Gerber and Green, 2012). This method retains the school within its randomised group but then models out the dilution effect of non-compliance by assuming that there exists a counterpart school in the control group, which because of randomisation should exist.

When?

At the analysis stage of the trial.

Why?

There are a number of common, but incorrect, methods of dealing with non-compliance. One is to simply exclude the non-compliant school from the analysis. However, this will introduce bias as the school that did not comply is likely to be different from the other schools and probably, due to randomisation, will have a counterpart in the control group that has not been removed. Therefore, this approach introduces bias and probably gives a bigger, biased, intervention estimate than the ITT method. Another biased method is to include the non-compliant school but put it in the control group in the analysis. This method exaggerates the bias that is experienced in the preceding approach. Finally, another approach that is used when paired randomisation has been implemented is to remove the non-compliant school plus its opposing pair in the control group. This method introduces selection bias as well and reduces statistical power.

Post-testing

Post tests should be administered either under exam-like conditions or by some blind to the group allocation. Test marking should be done by someone blind to group allocation.

Description

A post-test is the outcome measure. This will tell us whether or not the educational intervention has had a significant impact on children's academic achievement.

How?

All groups should be tested at the same time and under the same conditions. Having the test done under exam conditions is appropriate. Otherwise if the test involves one-to-one or small group assessment the assessor should be masked or blinded to the group membership of the children. Markers should be blind to the children's group membership.

When?

When all the trial interventions have been completed.

Why?

Tests should ideally be administered under exam-like conditions so that students/pupils cannot be helped. Tests should also be marked by someone who is 'blind' to the group allocation. If the marker is not blind they may consciously or unconsciously award higher scores to one group compared with the other. In educational trials it is rarely, if ever, possible to blind students to the nature of the intervention that they are receiving. However, it is often possible to ensure that test markers are blind to group allocation.

Primary outcome

The primary outcome is the outcome that determines whether or not an intervention is effective. It should be decided before the trial starts and needs to be stated in the trial registration document.

Description

The primary outcome is a single measure of outcome, which determines whether or not an intervention is effective.

How?

Discussion between the researchers, funders and educational experts about which is the most appropriate outcome measure for the trial should take place. The researchers, funders and educational experts also need to decide what difference is educationally 'worthwhile'. Generally an expensive and/or complex intervention needs a larger difference to be worthwhile compared with a simple and inexpensive intervention. Systematic reviews of previous, similar, studies will also drive the choice of outcome and the expected differences.

When?

The primary outcome should be determined before the trial commences. It should be recorded in the study protocol and in the registration document.

Why?

Sample size calculations are usually done on the premise that there is a single analysis done once at the end of the trial. Often more than one outcome is compared between the two groups. It is usually quite sensible to do this in order to gain a picture of the broader impact of the intervention, or to explore the mechanism of the effect, or to flag up avenues for further research. However, the results of these secondary analyses should be treated with some caution and should be used to support the results of the primary outcome, rather than to over-ride a statistically insignificant primary outcome. The problem with not defining the main outcome before analysis is that once the data are seen your choice of what should be the primary outcome is contaminated by the partial knowledge of the results. If you run a trial that measures 20 different outcomes there is a very high possibility that at least one of these will be statistically significant even if in reality there are no differences between the groups.

Pre-test equivalence

Baseline comparisons are commonly undertaken: this is inappropriate and may mislead the main outcome analysis. Comparisons of baseline characteristics of analysed groups may be helpful to spot selection bias due to attrition.

Description

During analysis it is common practice to compare, formally or informally, the groups on the basis of their baseline characteristics (e.g., number of boys per group; average age; average pre-test score).

How?

In the presence of sample attrition whether or not this may introduce bias can be explored by comparing the *analysed* groups at baseline. It is important not to confuse this with the baseline testing of *randomised* groups. The analysed groups are those who have post-test data, the randomised groups include those who were randomised whether or not they have post-test data.

When?

During the initial statistical analysis.

Why?

If there is attrition this may introduce bias. We can see if attrition has introduced bias in 'observable' and measured characteristics (e.g., age), but not on unmeasured variables. However, if there is no imbalance in observable variables and attrition is similar between groups this increases our confidence that there is no bias. On the other hand, the more common practice of establishing 'baseline' equivalence between the randomised groups is mistaken. This usually takes the form of comparing the baseline characteristics of the randomised participants. Commonly a table has columns for each group with pre-test scores, average ages, proportions of boys and girls etc. with t-test or chi squared test results denoting statistical significance levels. Unless we think that the randomisation has been manipulated this is pointless. Because we used randomisation we know that any differences between the groups are due to chance. Statistical significance or lack of significance should not be used as a guide to analysis. Invariably given sufficient numbers of baseline variables one or two will turn out to be 'statistically significant'. Assuming a robust randomisation process this is by chance: there is no true difference between the groups. Some may argue that baseline testing can be used to guide the statistical analysis of the outcome. For instance, it might be argued that if, by chance, we see statistically significantly more girls in one group than the other we should use gender to adjust the statistical analysis to take this into account. However, let us assume that gender is statistically significantly in imbalance but pre-test is not, to decide to adjust for gender, and not pre-test, would be a mistake. This is because pre-test is almost always a powerful

predictor of post-test results and should be used in an analysis of covariance whether or not it is in balance. We might wish to adjust for gender as well, but only if we think it is a powerful predictor, not because it is in imbalance at baseline. Furthermore, if a powerful variable is in imbalanced by a 'non-statistically significant' amount not adjusting will produce a biased estimate

Outcome analysis

Trial analysis comparing mean scores adjusted for baseline variables using regression based methods is usually the most efficient analysis. Analytical methods should be stated before data are seen.

Description

Usually we want to compare the means of our groups to see if any differences we observe are simply due to chance or are likely to be a 'true' difference (although chance as an explanation can never be entirely ruled out).

How?

There are a number of analytical strategies that can be adopted as well as intention to treat analysis. The choice of method should be stated in the protocol before the data are examined. For example, one may simply compare the two mean post-test scores using the two sample t-test. If we do not have a pre-test measure then this is a reasonable approach. However, if we have pre-test scores a regression type of analysis (e.g., Analysis of Covariance - ANCOVA) including pre-test as a variable in the analysis will improve the power of the study. Furthermore, we may include other variables that we think are important, such as age and gender, which may improve the power and precision of our study still further.

When?

During the statistical analysis at the end of the study.

Why?

A common approach to including pre-test scores in an analysis is to calculate 'change' scores and then compare the means of the change scores. These are simply calculated by subtracting the post-test means from the pre-test means. This approach has several limitations. First, you need to use similar tests for both pre and post- test for the method to work. Second, the approach is not as powerful as ANCOVA, if pre- and post-test correlation exceeds 0.5, which in educational trials is usually the case. Third, if there is some chance imbalance in pre-test scores then the method may amplify regression to the mean effects and not produce as valid a result as ANCOVA. Consequently, it is usually better in educational trials to use regression-based methods adjusting for powerful predictor variables and not calculate and use change scores. However, if change scores are used then including the pre-test score in a regression analysis with the change score as the dependent variable will produce the same result as an analysis just testing the difference in means using pre-test as a covariate. Note it is important to state the primary analysis method before the data are seen just as it is important to state the primary outcome.

Analysis of cluster randomised controlled trials

In cluster randomised trials the clustering must be taken into account in the analysis. Failure to do so leads to biased estimates of statistical significance. *Analyse how you randomise!*

Description

Usual statistical approaches assume independence of the each observation. When classes or schools are randomised children's outcomes are not independent of each other, they correlate. It is crucial this correlation is taken into account in the analysis.

How?

A simple approach is to treat each class or school as a single observation. Thus, for a class of 30 children we would calculate a class mean score (i.e., take the 30 scores of all the children and divide by 30 to calculate the mean). If we have 20 classes in our trial then a legitimate statistical method would be to do a t-test comparing the mean scores of the 10 classes in the control group with the 10 classes in the intervention group. The analysis of the trial is, therefore, of 20 class means not 600 individual children. Alternative methods such as multi-level modelling or the Huber-White method can adjust for pre-test measures and include individual children in the analysis. However, for large trials all the approaches tend to give similar results. For smaller trials with relatively few numbers of clusters complex methods, such as multi-level modelling, may not work. In these instances it is best to use the summary statistics method (i.e., comparing cluster level means).

When?

During the statistical analysis at the end of the study.

Why?

Failure to take into account clustering leads to a biased statistical significance test. Although the estimate of effectiveness is usually similar (although not always) the statistical significance value will be too small and may lead us to conclude, incorrectly, that an intervention is effective (or ineffective) when in reality there is no difference.

Clustering within individually randomised trials

Within individually randomised trials the analysis may still need to adjust for clustering if the intervention is delivered at a group level.

Description

When schools or classes are randomised there is clear clustering of outcomes. Clustering may still occur when individuals are randomised to a group activity. If, for example, children starting school with two entry classes we might randomise children to the two classes so that we could experiment with a different curriculum. Although the children's outcomes are not clustered at the start of the RCT they may well be at the end. This should be taken into account in the analysis

How?

Similar approaches as used in the analysis of cluster randomised controlled trials can be used.

When?

During the statistical analysis at the end of the study.

Why?

As with cluster randomised trials failure to take into account clustering can lead to biased estimates of statistical significance.

Secondary analyses

Secondary analyses as well as primary analyses need to be pre-specified to avoid data dredging and false positive results.

Description

Secondary analyses are additional analyses on outcomes on which the intervention may impact over and above the primary outcome measure. These may include subgroup analysis, where, for example we may look at the impact of the intervention in pre-specified subgroups (e.g., does the intervention work better for boys compared with girls?).

How?

Secondary analyses should be pre-specified to test a hypothesis *before* the data are initially analysed. They should be included in the trial protocol.

When?

During the statistical analysis at the end of the study.

Why?

It is perfectly acceptable, and often desirable, to test for a number of different outcomes. However, it is important that these are stated up front in the trial protocol to avoid the problem of 'data dredging'. Testing numerous different hypotheses will almost certainly generate statistically significant findings simply by chance, even where, in reality, there are no differences. Having a pre-specified analysis plan, which includes all of the analyses before the data are examined, is good practice and will reduce the problem of false positive findings.

Reporting uncertainty

Trial results should be reported using both exact p values and confidence intervals (usually 95% intervals).

Description

All estimates of effectiveness from a RCT are surrounded by sampling uncertainty. The larger the sample size the smaller is this uncertainty. Nevertheless, we will always be unsure of the exact intervention difference. It is important to characterise this uncertainty.

How?

We should report confidence intervals or Bayesian credibility intervals. These are usually reported as 95% confidence intervals. For example, if we see a difference of 10 points in a test and the 95% confidence interval is 5 to 15 points, we can be fairly confident that the true effect of our intervention falls within this range. Note, however, the most likely 'true' difference will be around 10 (it is unlikely to be exactly 10 points, however). As we move further away from our central estimate, the likelihood of the higher or lower values declines. Most standard computer packages will produce confidence intervals as part of their output.

When?

During the statistical analysis at the end of the study.

Why?

Although we may see a difference between the two groups at the end of our trial we would like to know if that difference is due to the play of chance or is a 'true' effect. It is important to know what the random variation around our estimated effect is so we can be sure we have ruled in or out an educationally important difference.

Sample size calculations

We need to calculate a sample size before we start the trial that will be large enough for us not to miss a modest but educationally important difference.

Description

When we plan a trial we undertake a sample size calculation to ensure that the trial is large enough to observe what we think is an educationally important difference. Conventionally a trial has either 80% or 90% power with 5% significance to observe a pre-specified difference.

How?

There are many freely available computer packages that can calculate sample sizes. One that has been specially designed for randomised trials in educational settings is the Optimal Design Software (http://sitemaker.umich.edu/group-based/optimal_design_software). As a rule of thumb, for 80% power and assuming a 5% significance level we would need 128 individuals randomised to see a difference of 0.5 standard deviations; 200 for 0.40; 356 for 0.30 and 500 for 0.20. If the trial is randomised by school or class we will need in the order of 4-8 times the number of individuals. However, this sample size can be reduced by up to 60-70% if there is a strong pre- and post-test correlation.

When?

At the design stage of the trial.

Why?

Many educational trials are too small and may miss educationally important differences. Undertaking a sample size calculation before the trial starts allows us to estimate the differences we could detect and whether we might be missing smaller but important effects.

Allocation ratios

Equal randomisation gives the most powerful trial when the constraint is the number of children or schools available. In the presence of resource constraints increasing the sample size and using unequal allocation favouring the resource unconstrained group increases power.

Description

In most trials schools and students are randomly assigned in equal numbers to the intervention and control groups. For a given total sample size this makes sense in that we usually gain the maximum statistical power when the group sizes are numerically equivalent. However, there are times when we might randomly allocate more schools or students to one group in preference to the other.

How?

When we ask a computer package to produce a randomisation series it can do this in a 1:1 ratio (i.e., one class, student into the control group and one in the intervention) or it can do other ratios (e.g., 3:2; 2:1; 3:1) where we may deliberately allocate more to one of the groups.

When?

At the design stage of the trial.

Why?

An important reason for using unequal allocation may be resource constraints. Here we use the term resources not simply to denote financial costs but other constraints such as availability of teaching staff or the supply of the intervention. Consider a mentoring programme, for example. Let us assume we have 20 teachers willing to mentor two students each. If we undertook a randomised trial of this using equal randomisation we could only include 80 students: 40 being mentored and 40 control students. However, we would get a more statistically powerful study if we recruited 120 students and randomly allocated 40 to be mentored and the remaining 80 to act as controls.

Pilot trials

Pilot trials test out the procedures of the main trial and their results can be integrated into the main trial. Feasibility trials may seek to change the intervention or outcomes and cannot be integrated into the main study.

Description

It is often useful to undertake a small pilot or feasibility study before we embark on the main trial. A pilot or feasibility study is not intended to answer the main question but to address the question of whether a large study can be conducted using the proposed trial design.

How?

We would recruit a small number of schools or students to test out the trial processes before embarking on the main trial.

When?

Before the main trial begins.

Why?

There are differences between a pilot and feasibility study although the terms 'pilot' and 'feasibility' are often used interchangeably. We might think of a pilot as the equivalent of the main trial in all respects except for its size. A pilot trial might look at the feasibility of recruitment of schools and students to the trial. Some pilot trials are internal pilots where we fully intend to run on into the main trial and the primary function of the pilot phase is to inform us about issues in the main trial, for example, whether we need to recruit 30 or 40. An external pilot is a small trial that stands alone and may inform the size of the main trial and its likely costs. Because we are not changing the intervention or the outcome we can combine the pilot phase results with the main trial. Sometimes, however, a pilot trial may change into feasibility trial. A feasibility trial is a small trial where we test the intervention and the suitability of the outcome measures. We may include qualitative or process evaluations in such a study so that the intervention can be changed or adapted or outcomes can be refined. Because of this it is unlikely that the results of a feasibility study can be combined with the main trial results.

Sample size calculations for pilot trials

Sample sizes of pilot trials should be approximately 9% of the sample size of the main trial or 30 individuals, whichever is the larger. If it is a pilot cluster trial then a minimum of 8 clusters is recommended.

Description

Sample size calculations for pilots can be tricky as the usual approach described above does not work because we are not looking for a difference between the groups (by definition the pilot will be too small).

How?

One approach would be to have a sample size where we could construct a one-sided 80% confidence interval that would exclude the educationally important difference if the point estimate was either zero or negative. Other simulations suggest we should not recruit fewer than 30 participants in total. If we are piloting a cluster randomised trial then ideally we suggest that at least 8 clusters should be randomised.

When?

At the design stage.

Why?

If the difference is zero or negative (i.e., favouring the control condition) then it is unlikely the main trial will find an educationally important difference. Eight clusters appear to be the minimum that can be used to undertake some basic analysis whilst controlling for clustering effects.

Balanced design

A balanced design might be helpful to control for the time and attention effects of the intervention and is efficient in that two interventions can be tested together.

Description

A balanced design is when we allocate the 'control' schools to an alternative intervention, which deals with the time and attention that the intervention schools are getting. For instance, if we wanted to test a new maths curriculum we might randomise the control schools to receive a novel literacy curriculum. This will deal with the issue of 'resentful demoralisation' where the control schools feel let down because they have not received any intervention. Furthermore, it will allow us to evaluate both the maths intervention and the literacy intervention at the same time.

How?

Choose a control intervention that will not 'spill over' and affect the other intervention's treatment effects.

When?

At the design stage.

Why?

Balanced designs are efficient and deal with the Hawthorne and preference problems of using an untreated control group. Care needs to be taken in choosing the control intervention as there may be spill over effects. For example, a physics intervention is likely to modify the impact of any maths intervention.

Factorial designs

The use of factorial trial designs is efficient. If an interaction between the two interventions is unlikely this allows 'two trials for the price of one' in sample size terms.

Description

The use of factorial trial designs is a method of evaluating two interventions in the same total sample size as a standard two armed randomised controlled trial. In the simplest factorial design we have four groups instead of two. Let us assume we want to evaluate the effect of phonics teaching versus a whole language approach *and* we want to evaluate delivery of both interventions either as a one to one or class delivery approach, we could use a factorial design to answer both these questions at the same time. In the table below we illustrate the design.

	One to one teaching	Class teaching	**Comparing phonics versus whole language teaching**
Phonics teaching	**Group A** Phonics delivered one to one	**Group C** Phonics delivered whole class	**Estimate total mean scores of A + C**
Whole language teaching	**Group B** Whole language delivered one to one	**Group D** Whole language delivered whole class	**Estimate total mean scores of B + D**
Comparing one to one versus class teaching	**Estimate total mean scores of A + B**	**Estimate total mean scores of C + D**	

How?

Schools or students are randomised into four or more groups instead of the usual two. In this example we are answering two questions: (1) Does phonics teaching improve outcomes more effectively than whole language teaching? and (2) Does one to one teaching improve outcomes more effectively than whole class teaching?

When?

At the design stage and when randomisation occurs.

Why?

For a factorial trial we essentially get two trials for the 'price of one' in terms of sample size. For the sample size calculation we estimate the smallest difference we want to detect between the two comparisons and then calculate the sample size for that. In terms of analysis we analyse as two separate trials (as previously described). However, we would normally test for an interaction in case the intervention effects are not additive. In a factorial trial we assume that the interventions *are* additive, which means that they work just as well in the presence or absence of the other intervention. Sometimes, however, this may not be the case. Let us assume, for example, that phonics teaching improves outcomes more effectively when it is taught using a one to one approach rather than using a whole class approach. Or let us assume that one of the methods of teaching is more or less effective using one-to-one approach or whole class approach. In this instance we have an interaction, which can damage the power of the trial and its interpretation. Consequently, when planning a factorial trial, we need to be reasonably confident that there is not likely to be an interaction between the interventions. Nevertheless, it is usually good practice to test for an interaction in the analysis. However, even if there is an interaction it is unlikely that it will be statistically significant at conventional levels of significance as the power of an interaction test is low. Usually, however, we try to compensate for this by using a higher level of statistical significance (e.g., $p = 0.10$ instead of $p = 0.05$, for the interaction test). It is important to note here that a 2 x 2 factorial is *not* a four armed trial, as the analysis and sample size calculations are quite different between the two types of design. If we were doing a four armed trial we would answer a different set of questions. Because we are comparing single cells with another single cell we lose power so we need to increase our sample size in a four armed trial compared with a factorial trial Also because we are undertaking multiple comparisons we need to adjust the sample size to take this into account (i.e., increase it yet further).

Split plot

A split plot trial is a type of factorial trial where schools or classes are randomised and then individual students are also randomised.

Description

A split plot design is a form of factorial study. In this instance we might randomise schools to a whole school intervention or whole school control condition and then randomise individual children within the cluster conditions to other interventions or control conditions. For instance, we might randomise the school to receive extra investment in computer technology and then randomise individual children to receive exposure to a particular software programme to aid numeracy acquisition. A variant on this design is a *partial split plot* whereby we randomise at the school or class level and then randomise individual children in the interventions schools to get an enhanced version of the intervention. For example, schools might implement a new numeracy curriculum, but children within the intervention schools are then randomised to receive the intervention either at the whole class level or in small groups.

How?

Schools are randomised then individual students within the schools are randomised.

When?

At the design stage and when randomisation occurs.

Why?

Using a split plot design an ascertain whether the curriculum has positive effects on school level outcomes but also enables us to assess within the intervention schools whether delivery via small groups is more effective than delivery at the class level.

Stepped wedge

Stepped wedge trials are a type of cluster trial. They can be useful in evaluating interventions that for political or other reasons have to be rolled out nationally. They also can be useful when there are only a few clusters as they can increase statistical power.

Description

The stepped wedge design is a form of cluster trial. In the standard cluster trial we would recruit our schools, pre-test the children and then randomise them into, say, two groups of 20 schools. We would then offer the intervention to half the schools and give post-tests at the end of the study. Thus, we would have two groups and two tests (pre and post-tests). In a stepped wedge design we would take our 40 schools, pre-test the children as in a cluster trial and then randomly allocate some of the schools to the intervention.

How?

Let assume we have a four period stepped wedge trial with 40 schools. In this instance we would randomise 10 schools to get the intervention, and then, after, say, the first term, we would do post-testing. At this point the trial is the same as a cluster trial, albeit with unequal allocation. However, the trial does not finish at this point: from the remaining 30 control schools we randomise another 10 schools to receive the intervention. We follow up all 40 schools for another post-test, after which we are left with 20 control schools. We then randomly select another 10, intervene and post-test all, and then finally the remaining 10 schools are offered the intervention.

When?

At the design stage and when randomisation occurs.

Why?

There are advantages and disadvantages to using a stepped wedge design. First, if for some reason we need to implement the intervention to all schools this is done as part of the intervention. This may reduce resentful demoralisation as all the schools know they will get the intervention eventually. A disadvantage, however, is that if we find that the intervention has a harmful effect on educational outcomes by the time we know this it is too late for the control schools. Second, because we have repeated testing, the design may give us extra statistical power for the same sample size compared with a two armed cluster trial. Conversely, the design does involve more post-tests and this may increase the costs of the study.

Reporting RCTs and the use of the CONSORT statement

All RCTs should be reported according to the CONSORT criteria.

Description

In the mid-1990s a group of health care RCT methodologists and medical journal editors got together and published the CONSORT statement. This statement consisted of a number of items that in future many health care research journal editors would insist that trial publications should report. For example, a full description of the randomisation method should be reported in detail to allow the reader to judge the likelihood of allocation subversion; the rationale behind the sample size should be described, as should the primary outcome and the statistical analysis method. Since its adoption by most health care journals many journal editors outside of health care have adopted the guidance. Indeed, many psychology journals and educational journals state that they want RCTs to conform to CONSORT standards. The original statement has been adapted to fit with educational RCTs.

How?

Go on the CONSORT website for detailed guidance on how to report a trial using its guidance (http://www.consort-statement.org/). A key feature of the CONSORT guidance is the CONSORT flow diagram. The use of a flow diagram aids understanding of the trial and what happened to potential participants, those who refused randomisation, and actual participants after randomisation

When?

At the design and writing up stages of the trial.

Why?

It is important that the trial is designed using CONSORT as a template so that its reporting can be follow the criteria. Although CONSORT was designed to improve trial reporting it can also be used to help design a trial. If researchers realise that their trial will be reported according to CONSORT then it makes sense to ensure that the design of the study will enable the report to fit in with what the guidance asks for. Consequently, using CONSORT to inform trial design should increase the quality of the trial.

Trial Checklist

	Action	Rationale
1	Has a sample size calculation been undertaken?	To ensure trial is not too small.
2	Has a primary outcome variable been specified?	To avoid data driven outcome selection
3	Has trial been registered?	To avoid publication bias.
4	Have eligible children/students been identified within the school before randomisation?	To avoid recruitment bias.
4	Has pre-test been performed before randomisation?	To avoid biased pre-test results.
5	Has randomisation been done by an independent third party (concealed allocation)?	To avoid allocation mishaps/subversion.
6	Have post-tests/outcomes been undertaken independently and marked blindly?	To avoid ascertainment bias.
7	Has intention to treat analysis been done?	To avoid selection bias being introduced.
8	Has analysis taken clustering into account?	To avoid biased standard errors.
9	Have the trial results been published?	To avoid publication bias.

Common questions in trial design, conduct and analysis

The following questions are all either derived from errors found in published RCTs of educational interventions or actual questions asked of the authors by other researchers and students.

Question: "I can't conceal the random allocation because the teachers and the students know what intervention they will be getting".

Answer: "You can always conceal the allocation. You are confusing this with blinding or masking of the delivery or receipt of the intervention, which is rarely possible in educational research. Concealed allocation means the teachers and children do not know in advance of the randomisation process which intervention they will be allocated to".

Question: "One of the schools has expressed a preference for the intervention. Shall I go ahead and recruit and randomise in the hope they will get it?"

Answer: "No, it is better not to recruit that school unless you can convince them to accept the chance that they get the control intervention. If not, do not recruit this school, otherwise they will probably drop-out if allocated to the control group".

Question: "I have randomised 100 students into the trial, unfortunately 10 in the intervention group no longer want to take part. Can I replace them from my waiting list?"

Answer: "No you can't do this as doing so will introduce selection bias. The groups were balanced at baseline due to the randomisation; replacing drop outs from elsewhere cannot deal with this problem. You should try and keep the drop-outs in for the post-tests and adjust your analysis using Complier Average Causal Effect analysis".

Question: "If I randomised 120 students instead of 100 students and have 10 'waiting list' students for each group can I use them to replace those who drop out?"

Answer: "No, it is exactly the same problem as taking students off a waiting list; it does not matter that they were randomised within the waiting list or not. This is because those who drop out are likely to be 'different' from the average student in the trial and the replacement student is not likely to have the same characteristics of the student who left the trial".

Question: "It is really easy to allocate people by the last digit on their date of birth or by their UPN or National Insurance number (odd numbers get the intervention even get the control). Can I do this instead of randomisation?"

Answer: "No, this approach allows allocation mishap as people know who should get which intervention and so can preferentially exclude students they don't think will respond best".

Question: "The allocation that I was given for the 20th randomisation shows it goes into the group that already has 10 schools. This must be a mistake; shall I override the randomisation and put it into the smaller group of 9 schools to ensure I have equal numbers?"

Answer: "No, randomisation rarely produces identical group sizes. It is best not to interfere with the allocation schedule: accept what the independent allocator has given".

Question: "We are going to randomised 40 schools we have data to allow us to randomise the first 20 but are still waiting to for stratifying data for the last 20. We need to start the intervention now in some schools can we randomise the first group and then randomise the second 20 when they have returned their data? Or do we have to wait until all 40 schools are ready?"

Answer: "Yes it is perfectly acceptable to randomised the first 20 and then add the remaining schools when they are ready. If you use the minimisation technique to restrict your allocation this will automatically take into account the balance of the existing schools allocated.

Question: "Is it best to have a single teacher deliver the intervention or multiple teachers?"

Answer: "It is best to have as many teachers as possible to deliver the intervention to ensure generalizability. If you use a single teacher you will not be able to disentangle the effects of the teacher from the effects of the intervention."

Question: "I am testing a maths and literacy intervention; can I use the maths schools as the control schools for the literacy intervention and vice versa?"

Answer: "Yes, you can if you are confident that effects of the literacy intervention will not spill over to the maths and vice versa. This is an efficient design".

Question: "Some of the students didn't turn up for any of the intervention at all. Should I not bother to include them in follow-up and analysis?"

Answer: "It is crucial that you retain these students in the post-test; if you do not follow them up you cannot do intention to treat analysis and you are likely to have a biased result".

Question: "One of the teachers randomised to the intervention did not deliver it properly to her students. Shall I exclude her class from follow-up and analysis?"

Answer: "Absolutely not. The students should be tested and their results included in the analysis, otherwise you will introduce bias".

Question: "We have randomised individual students within schools. We have found that in some schools compliance with the intervention is very low. Shall we exclude those schools and only include those with high compliance because after all we are using individual randomisation and each school has equal numbers of control and intervention students so we will be excluding both sets of students, which will avoid selection bias."

Answer: "Whilst this appears to be a good solution, it is not as selection bias can still be introduced. It is better to include all of the schools as this will avoid bias. Also it will produce a more policy relevant answer as inevitably there will be some schools that have poor compliance/implementation."

Question: "It is easier, and cheaper, for the teachers to give the post-tests to the controls and use the researchers to give the post-tests to the intervention group. Can I do this?"

Answer: "No, this will potentially introduce bias as students may react differently when given the test by their teachers rather than by the researchers. Both groups need to be tested in the same way."

Question: "I have 90 post-test results available but not all of them have a pre-test value – should I exclude those without the pre-test in my analysis?"

Answer: "You should do at least one analysis just comparing the mean values of all the post-test scores as this estimate is the one least likely to be prone to bias. You could also do a complete case analysis, which is to only analyse those with both pre- and post-test data. You could also impute the missing values. If all the methods give similar results you will be confident that the missing pre-test data are unimportant. Different results, however, would mean the post-test only analysis is more conservative and less likely to be biased."

Question: "I am worried that the randomisation may not have 'worked' properly; shall I check by comparing pre-test and other baseline variables using statistical testing?"

Answer: "If you think there may be a problem in terms of your randomisation system, baseline testing may not help you. It is entirely possible for the baseline testing to show no statistically significant differences between groups and yet the randomisation is compromised. It is far better to ensure you have a robust system in place to deliver proper unbiased randomisation. If you have a robust system then baseline testing is irrelevant and potentially misleading as the null is true and you might be guided by inappropriate testing to adjust for variables that are in imbalance but not for stronger covariates that are better balanced".

Question: "I have two schools only; can I randomise these and then compare the mean scores statistically".

Answer: "No, if you only have two classes or schools this is equivalent to having two students. Randomisation in those circumstances cannot balance out any characteristics of the school or class that might affect outcome. You need cluster replication of at least 4 schools per group to do this".

Question: "What is the minimum number of schools or classes I need?"

Answer: "This depends on the difference you are looking for and the analytical technique proposed. Some suggest that a minimum of 4 schools per group is necessary and others say 7. Certain statistical methods, such hierarchical linear modelling (HLM) (needed to adjust for clustering) are not robust if the group size is less than 15 schools or classes per group".

Question: "I only have 14 schools in my trial (7 in each group). This is too small to use HLM; shall I use statistical approaches that ignore clustering?"

Answer: "No, if you do this you will get a biased estimate of the standard error. You will mislead yourself and the reader into believing that something is effective, or ineffective, when any difference may be due to chance. Cluster level mean differences is a conservative approach but takes clustering into account and can be used in small trials".

Question: "I have found a modest difference, which is statistically significant, between the groups, but when I looked at the results for boys only I found a much larger, statistically significant, effect of the intervention for them. But when I looked at just the girls the difference was smaller and not statistically significant. Shall I recommend that only boys are offered the intervention?"

Answer: "No, it is likely that the difference between the sub-groups is by chance. Your main finding, of a modest difference for all children, is most likely the correct finding. You could undertake a statistical test looking at the interaction between gender and post-test. If this is statistically significant then this would suggest that further research in another trial looking specifically at this issue is warranted. In future trials you should ensure you pre-specify subgroup analyses or interaction tests."

Question: "In my RCT of different types of maths teaching I found a statistically significant interaction between the results of the new maths teaching and the gender of the teacher (i.e., when the teacher is a male the results are better than when the teacher is female). Should I recommend that this maths curriculum is delivered by male teachers?"

Answer: "No, although your interaction is significant this is of interest for further research. Because you did not randomise the teachers, only the curriculum, it is likely that the interaction is due to selection bias rather than a 'real' effect. A future study could randomise the new curriculum to be delivered by male or female teachers: this would give a robust answer".

Question: "My 'negative' trial has been rejected by a number of journals as having 'uninteresting' results. Should I give up?"

Answer: "Absolutely not. Keep trying the journals and if necessary post the paper on a website. It is crucial all trials, whatever their results, are reported in the public domain."

Suggested Further Reading

Gerber AS, Green DP. (2012) Field Experiments: Design, analysis and interpretation. WW Norton and Company, New York.

Haynes L, Service O, Goldacre B, Torgerson D. (2008) Test, Learn, Adapt. Developing Public Policy with Randomised Controlled Trials. Cabinet Office Behavioural Insights Team.

Shadish WR, Cook TD, Campbell TD. (2002) Experimental and quasi-experimental designs for generalized causal inference. Houghton-Mifflin Co, Boston.

Torgerson DJ, Torgerson CJ. (2008) Designing Randomised Trials in Health Education and the Social Sciences. Palgrave McMillan, Basingstoke.

References

Boruch RF. (1997) Randomized Experiments for Planning and Evaluation: A Practical Approach. Applied Social Research Methods Series 44, Age Publications London.

Chalmers I, Glasziou P, Godlee F. (2013) All trials must be registered and the results published. British Medical Journal 346:f105 doi: 10.1136/bmj.f105

Cocks K, Torgerson DJ. (2013) Sample size calculations for pilot randomised trials: a confidence interval approach. Journal of Clinical Epidemiology 66; 197-201

Gerber AS, Green DP. (2012) Field Experiments: Design, analysis and interpretation. WW Norton and Company, New York.

Goldacre B. (2008) Bad Pharma Fourth Estate, London.

Haynes L, Service O, Goldacre B, Torgerson D. (2008) Test, Learn, Adapt. Developing Public Policy with Randomised Controlled Trials. Cabinet Office Behavioural Insights Team.

Hewitt C, Hahn S, Torgerson DJ, Watson J, Bland JM. (2005) Adequacy and reporting of allocation concealment: review of recent trials published in four general medical journals. British Medical Journal 330, 1057-1058.

Lindquist EF. (1940) Statistical analysis in educational research. Boston, Houghton Mifflin.

Puffer S, Torgerson DJ, Watson J. (2003) Evidence for risk of bias in cluster randomised trials: a review of recent trials published in three general medical journals. British Medical Journal 327, 785

Schulz KF, Chalmers I, Hayes RJ, Altman DG. (1995) Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. Journal of the Americal Medical Association 273, 408-412.

Torgerson CJ. (2003) Systematic Reviews Continuum: London

Torgerson CJ, Torgerson DJ, Birks YF, Porthouse J. (2005) A comparison of randomised controlled trials in health and education. British Educational Research Journal 31,761-785.

Torgerson DJ, Torgerson CJ. (2008) Designing Randomised Trials in Health Education and the Social Sciences. Palgrave McMillan, Basingstoke.

Torgerson CJ, Wiggins A, Torgerson DJ, Ainsworth H, Hewitt C. Every Child Counts: testing policy effectiveness using a randomised controlled trial, designed, conducted and reported to CONSORT standards (2013) Research in Mathematics Education, 15, 141-153.

.