Issue 5, March 2007

THE UNIVERSITY of York

Trials in Public Policy

Inside this issue:

1) Second Annual Conference

Randomised Controlled Trials in the Social Sciences: The Way Forward

12th—14th September 2007

Following the success of the inaugural conference — Randomised Controlled Trials (RCTs) in the social sciences: Challenges and prospects, we are pleased to announce the second annual conference addressing the theme RCTs in the social sciences: The way forward.

This conference aims to:

- Promote and develop knowledge of Randomised Controlled Trials (RCTs) in the Social Sciences
- Facilitate sharing and collaboration between academic disciplines in this area
- Contribute to the understanding and conduct of evidence-informed policy and practice

Call for Abstracts

Abstracts are now invited for refereed oral and poster presentations relevant to:

- New thinking
- Ideas to advance use of trials evidence
- Examples of new trials
- Methods pieces

Abstracts should:

- Follow a structured format – such as Introduction, Methods, Results, Discussion for research studies
- Be no longer than 250 words, excluding references
- Include no more than three references

2) RDI Trials in Public Policy: 2nd Trainers' Event

18th April 2007

Encouraging the understanding of 'fair tests' in all areas of public policy, including criminal justice, housing, education, health promotion, public finance, and civic engagement.

Following the successful event for research methods teachers, held by the RDI Trials in Public Policy project in York last May, I am writing to let you know that the second such event will take part at The University of Teesside on April 18th, 2007. We hope that everyone who came to the first event will be able to attend. We will also welcome anyone unable to attend last time who has an interest in social science research methods teaching and its improvement. Note that the event and lunch are free, and that the project will refund the travel expenses of those previously registered.

The day will be interactive throughout, but loosely structured around the achievement of three objectives - curricular development in relation to teaching about trials, the creation and adoption of web-based resources for teaching about trials, and the improvement of the 'edited collection' on our website as a resource for teaching about trials (http://trials-pp.co.uk/). These three themes will be addressed in consecutive workshops led initially by Stephen Gorard and Carole Torgerson. It is important that, as a group, we make considerable progress with all three themes.

Participants are invited to bring curriculum, assessment or other teaching materials relating to these workshops for discussion/development.

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

RDI Researcher Development Initiative

We would like to introduce the first in a series of summaries from last years conference, Randomised Controlled Trials in the Social Sciences: Challenges and Prospects .

3) Regression Discontinuity
Tomas Cook

Cook and Wong (2007) discuss the use of the regression discontinuity design (RD) and ague for its wider use when RCTs are not possible. The RCT is the most efficient method of establishing a causal relationship between two interventions; however, it is not the only approach. Theoretically the RD design can produce as unbiased an estimate of effect as a RCT. There are many instances where it is not possible to undertake a RCT, due often to political or practical rather than ethical reasons, and other designs to establish a causal relationship have to be sought. The RCT and RD have a common feature: participants are selected for an intervention using a known and measurable variable. In a RCT this is achieved by random allocation. After the random allocation we know the random allocation status of each participant, whereas with RD we can achieve this measuring some characteristic of a participant and determining group allocation on this measurement. As an example, Cook and Wong discuss an intervention aimed at children who are gifted and talented. We might measure a child's IQ and offer an intervention to those who score 140 or higher and not offer it to those who score lower. Or we might offer a school or university scholarship to those who are from families with incomes below a certain threshold and not offer the scholarship to those above the threshold. University students might be offered increased support if they have an A level point score below a threshold and not if they have one above. To assess the effectiveness of an intervention we would then follow-up the whole cohort of participants: those above and those below the threshold, and essentially plot their outcomes against their baseline score which determined their group allocation. If the intervention has an effect on outcomes we would expect to see a break or a 'discontinuity' at the point where the assignment on the basis of the allocation score was made. If there was no effect there would be no break in the regression line. Mathematically this approach has been proven to be as unbiased, if implemented properly, as a properly implemented RCT. Nevertheless in the 'real' world few RCTs or RDs can be implemented perfectly. Consequently it is important to test empirically whether or not the RD design as implemented produces similar results to a real world RCT. After introducing the RD design and summarising the history of its development Cook and Wong then go on to produce an overview of 3 methodological studies where the outcomes of RD studies were compared with similar RCTs. All of these methodological experiments had some limitations; nevertheless, there was broad concordance between the RD and the RCT, thus giving support to the notion that where an RCT is either not feasible or practical then a RD should be the quasi-experimental design of choice. The RD design addresses some important limitations of the RCT. For those who are at the extreme of a distribution and it is thought to be unethical or infeasible to withhold an intervention from a random proportion of the population then the RD approach can allow us to evaluate the intervention in a robust manner. The drawbacks of the RD design are mainly twofold: first the cut point determining the allocation needs to be respected otherwise the design becomes 'fuzzy' and loses power; second, even when there is a sharp break the design is at least 2.75 times less powerful than the RCT. On the other hand power might be increased by the ability of the RD method to include a larger cohort in the study than an RCT. The RD design is an elegant solution to the problem of not being able to undertake an RCT; however, it is not used widely partly due to lack of dissemination and research support of this method. When a RCT is not possible researchers should automatically consider using a regression discontinuity design and only choose other quasi-experimental methods if this is not possible.

Reference for further reading:

Empirical Tests of the Validity of the Regression Discontinuity Design: Implications for its Theory and its Use in Research Practice ' Thomas D. Cook and Vivian C. Wong, Northwestern University.

What follows is an exploratory discussion about the meaning of the some of the terminology used in trials.

4) What is a Randomised Controlled Trial (RCT)?
Paul Marchant

The key is in the name. The word **Trial**; says we want to test something to see if it works. In the present context this might be in the field of education or in that of criminal justice. The trial might be to see either (1) if a particular method of aiding learning produces good results or (2) whether a method of reducing offending is effective. Of course many trials are conducted in the field of healthcare to see what constitutes effective treatment of ill-health. In fact the word '**treatment**' is the term which tends to be used for the method under examination in the language relating to trials, whatever the subject area.

The word **Controlled** says we compare the new method, in which we are interested, with another method of doing the same thing. That is so we can make comparisons of the effectiveness of the methods. Comparison is done is because the effect of temporal change can be eliminated. For example in a trial of a criminal justice intervention, if offending tends to reduce as criminals age, then it might appear that treatment for offending has worked when in fact it is nothing more than an ageing effect. Similarly the passage of time may have important bearing in education. By comparing the methods of achieving the same result, the temporal effect cancels. It is possible to compare several methods in one trial, but for simplicity at the outset we will just consider two treatments.

The word **Randomised** says that the allocation of methods under examination i.e. the treatments, is done at random. This means that it is only chance which determines which individuals get which treatment. This ensures that just after the allocation has been done, the groups are statistically equivalent. This is because the groups comprise random samples from the **population**, i.e. the wider group, from which the individuals involved in the trial were selected. The only difference the groups experience is that of the treatment they receive. Therefore if the groups are no longer statistically equivalent at a later point then this can be attributed to the effect of the difference in the treatment received. For example, if there is evidence that the mean of a measurement you are interested in differs between groups, by more than might be expected by chance, then it suggests that the treatments have had different effects. In fact, the difference in the treatment effect in the population is estimated by a statistical method, appropriate for the situation of the particular trial. It is inevitable that the estimated difference of treatment effect will be subject to uncertainty because our sample is not the whole population. Therefore the estimated treatment difference is usually presented in the form of a **confidence interval**, which gives a likely range for the true population treatment difference.

An example

Let us consider an example in order to introduce some terminology. Imagine a study in education in which we wish to compare two treatments (also known as **intervention**s) in terms of the benefit to students' learning. One treatment might be that of giving some additional computer aided learning whereas the other is that of giving additional printed materials. We envisage allocating individuals to receive one of the treatments at random. The allocation may be done so that the probabilities of entering either group are equal, but need not be. The individuals allocated are termed **unit**s. After the students have been exposed to the interventions they are tested for knowledge. The score on the test is called the **outcome measure** or **endpoint** and it is by this that the extent of learning is assessed. It is the comparison, by an appropriate statistical method, of the outcome in the different **arm**s of the trial, that the differential effect of the interventions is estimated. This is to enable us to see if one of the interventions is better than the other, as we use the data from our sample of students to infer back to the population; the wider group of students from which our sample comes.

Note the statistical methods used should be specified in the statistical analysis plan within the protocol, which is the description on the conduct of the trial, written before the trial starts.

Note it can be useful to utilise in the analysis, measurements taken soon after randomising, at the start of the trial. Such measurements are known as **baseline** values.

The kind of RCT design described above is known as a **parallel group** RCT as one envisages the research subjects travelling through the arms of the experiment in parallel.

Other kinds of designs, such as that of a cross-over trial, can be used in appropriate circumstances. More often however in educational trials yet another design can be preferable to that described above.

This is to randomly allocate whole teaching groups, classes, of students to receive the same one intervention and other classes to get the other. The unit of allocation here is class rather than individual student. Such designs are known as **cluster randomised trials**. (A synonym is '**group randomised trial**'.) It is likely however in such a case that measurements will be made on individual students. The matter of randomising at one **level** (class) and measuring at another lower level (student), leads to complications in the analysis. Therefore for the moment we will continue to think about parallel group trials where allocation and measurement is done at the same level, as in the individually randomised trial first mentioned above.

The RCTs give best evidence for treatment effects.

The beauty of allocation by chance is that factors which will influence the outcome, apart from the treatment given, are distributed equally between groups, on average. This includes things which we know influence matters, e.g. in an education trial students' past academic performance, but also more importantly not only things of which we have only poor knowledge but things we have never dreamed could be important.

RCTs give comparison of like with like so that the effect of treatment difference can be estimated. The analysis of the data follows logically from the design. In other types of study, e.g. in purely observational ones so that no random allocation applies, one must make allowances for confounding variables and this involves subjectivity of what to include and how to make the allowances. Such an approach can lead to **bias**, whereas the RCT design gives unbiased results. Bias means that the answer found is likely to be systematically different from the true answer. That is the estimate of the differential effect of the interventions will be either too big or too small, on average.

In trials we are making comparisons between different treatments. Sometimes, particularly in medical research, a dummy treatment (placebo, e.g. a pill of an inert substance) is used. Often the standard treatment, i.e. the one which is usually applied, is given in the control arm, as the comparator for some new treatment which it is hoped will be superior. In this case the control is known as TAU (**Treatment As Usual**).

Sample size

At the outset it is important to calculate what your RCT experiment is capable of delivering. For example, if one has an idea of the difference of effect between the two treatments one ought to ask whether this difference can reliably be detected with the sample size envisaged. It is sound practice that one calculates the sample size required for a trial so that an anticipated effect difference between treatments can reasonably be found. A sample size which is too small will not be able to detect the difference sought, although in some circumstances it may be sensible and ethical to run such a small trial. On the other hand a trial which is larger than necessary will use too much in the way of resources, not least the time and trouble of the research subjects involved. (Sometimes such a sample size calculation is called a **power calculation**, involving as it does the concept of **statistical power**, the ability of the trial to detect a treatment difference which truly exists.)

5) Concerns about sampling: Some comments on 'What is a randomised controlled trial?'

Stephen Gorard

It is extremely helpful to have what sometimes appear to be complex ideas explained simply and lucidly, as Paul Marchant has achieved here. Readers new to this kind of research will discover the meaning of many of the technical terms used in trials – terms that are often useful shorthand for researchers but sometimes off-putting for everyone else. A problem with the usual style of academic writing, for me, is that I often end up not knowing whether I disagree with something or merely do not understand it. This, in itself, is a good reason for us all to write and argue more clearly. It is how we learn, and so build the research-capacity 'capital' in our community. So, it was also useful for me to read this brief and simple account because it raises a couple of issues where I may disagree with Paul (and others).

First, I want to clarify my view of the sampling procedure in a trial, and its implications for analysis. For Paul, the two (or more) groups of individuals (or cases) used for a trial represent '…random samples from the population, i.e. the wider group, from which the individuals involved in the trial were selected'. He makes effectively the same point later when discussing the need for confidence intervals, again when explaining the need for a statistical treatment such as a significance test. '…We use the data from our sample of students to infer back to the population; the wider group of students from which our sample comes'.

The kinds of techniques that Paul mentions, such as statistical treatments, were indeed designed to help estimate the extent to which the measured difference in outcomes between two groups can be attributed to their earlier randomisation. But the population in this kind of trial is not some group larger than the cases participating in the trial unless the cases were selected for participation at random as well as allocated to treatments at random. This is very unlikely, and I have never encountered it. Because, in many trials, the cases taking part are volunteers they cannot be deemed statistically representative of some larger population. The population, in a normal trial, is then the total number of cases in the trial. Each group (or arm) is intended to be an equivalent random sample from that clearly defined population. So, in my opinion, all of the population takes part in the trial, and no further generalisation is possible. What does this mean for techniques such as confidence intervals and tests of significance?

These techniques can still be used to try and estimate the extent to which any outcome difference is attributable merely to the random allocation of cases to groups. However, they do not achieve this in the way that many might imagine. Although I outline the role of significance tests here, it is important to realise that equivalent limitations apply also to confidence intervals, and indeed to any technique predicated on sampling theory. The p-value from a significance test applied to the outcomes from two groups in a trial can only be calculated on the strict assumption that there is no difference due to the treatment of the two groups. Given this lack of difference, 'p' is then the probability of finding a difference in the group scores, at least as extreme as the one found, simply due to the random allocation process. Thus, the p-value cannot be used, directly, to estimate whether there is any non-random difference in the scores between the groups.

As soon as we allow that the two treatments might actually differ then we no longer have that p-value to work with, because it is predicated on their being *no difference*. Many commentators confuse the probability of finding data this extreme if there is no difference between treatments (the p-value) with the probability that there really is a difference between treatments given data this extreme (which is what we might really want to know). The former can be used to help estimate the latter, but this involves Bayes' Theorem, and is almost never done. The key point is that there is no standard statistical method that can be used to decide whether any observed difference between groups is large enough to merit substantive explanation.

In his example, Paul raises the common situation where the measurements are taken for individuals (such as pupil test scores) while the randomisation to treatments is for groups of individuals (such as schools). He claims that this 'leads to complications in the analysis'. This is not necessarily so. Any analysis should be conducted for the units that are randomly allocated to the treatments (e.g. the schools). The fact that there are different elements contributing to the overall score for schools is no more relevant than the fact that the score for any individual might be the sum of a number of different test items. If we wish to analyse individuals then, in my opinion, we need to randomise individuals. If this is not possible for any reason then this clearly means that we cannot analyse at the level of individuals using any techniques derived from sampling theory, with its explicit assumption that cases are randomly allocated or selected.

My final, related, area of mild disagreement lies with what appears to be the circular nature of the standard procedure for calculating the sample size. If our research is not important then we should not be wasting resources doing it. If we have no reason to believe that an intervention is considered plausible (preferably by people other than the evaluator), then we should not trial it. Thus, we should only go to the expense and bother of a trial where it has a fair chance of showing that an intervention works, and where the impact of the intervention would also be important or useful. We would then want this research to help influence policy or practice (i.e. to make a difference in the real world). In these circumstances we would want our findings to be as secure as possible, and this would demand, among other things, the largest sample size that resources would allow. The maximum sample size for research is more likely to be determined by resources than by sample size calculations. Thus, I believe, that calculations of the kind proposed by Paul should be used, at best, to help decide whether it is feasible to test an intervention with the maximum sample size possible. But even in this they fail, and for the same reasons as significance tests do.

Power calculations relate the size of the actual difference between the treatment groups, the variability of the outcomes scores in the population of interest, and the significance level sought, to find the minimum sample size required to detect a genuine difference between groups while trying to minimise the danger of being misled by a false difference. Each of these four items depends on the others. In general, we are more likely to find any 'genuine' difference between our groups if the treatment effect is strong, the variation in scores is low, our significance level is high, and our sample is strong. We do not know the first two of these values, otherwise we should not be doing the research. We can, therefore, only do the power calculation by imagining.

that the effect and variability are of a particular size and then seeing how the other numbers come out. But for any level of significance, picking the effect size and variation is effectively the same as picking the sample size. The ensuing 'calculation' is somewhat sleight-of-hand for an experienced researcher. Since these two values are not under our control as researchers anyway, we may as well simply go for the largest possible sample size, because this is just about the only uncomplicated way of improving the quality of our study. In the theory of power calculations widely used by Paul and others, the only other element that is under our control is the significance level. Setting it higher increases our chances of finding the difference we seek, so the theory goes, but also increase the chances of a false result. Therefore, we should set the level according to the dangers or costs of being wrong either way, when making our subsequent decision about whether the intervention has had any impact. But this theory must be false because it rests on contradictory premises. As discussed above, p-values (the basis of significance tests) can only exist if there is no difference between the treatments. To attempt to relate p-values and the size of the real difference between the treatments, as power calculations purport to do, is to make a clear logical error.

Further contributions are welcome with a view to carry this discussion forward in further issues. For more information please contact : educ-trials-pp@york.ac.uk

RCT Help Line

If you have a query or would like help or advice on any aspect of designing, running or evaluating
randomised controlled trials, please contact us. Where appropriate, a member of the project will be happy to visit the site to provide personal assistance.

Contact Us:
Tel: 01904 433466 or
Email: educ-trials-pp@york.ac.uk

We want to encourage more people to be involved in face-to-face events, and in virtual participation, from all areas of public policy. In particular, we want to hear from national, regional and local policy-makers and practitioners who do or could use evidence from rigorous evaluations in their fields. And from research methods trainers, struggling with the place of trials methods in their courses. The first two events were in York, but we are happy to hold or help organise events wherever they are wanted. Please contact us with your comments and suggestions.

The RDI Trials Project Administrator
Department of Educational Studies
University of York
Heslington
York
YO10 5DD

Phone: 01904 433466
Fax: 01904 433459
Email: educ-trials-pp@york.ac.uk