CHAPTER SEVEN

RANDOMIZED CONTROLLED TRIALS

Carole J. Torgerson, David J. Torgerson, Celia A. Taylor

When policymakers require evidence of "what works?" in public policymaking, this often involves the use of quasi-experimental evaluations, which have been described in detail by Gary Henry in Chapter Six of this book. As Henry notes, randomized controlled trials (RCTs) are generally acknowledged by methodologists to be the gold-standard method in evaluation research. This is so because RCTs enable causal inferences to be derived through the observation of unbiased estimates of effect (Shadish, Cook, and Campbell, 2002), provided they are rigorous in their design and execution. The randomized trial is one of the key methodological breakthroughs for program evaluation that has occurred in the last one hundred years. Therefore, when an RCT is possible, this design should generally be used in preference to alternative evaluative approaches.

The RCT is a simple concept. If participants are allocated to one or more groups *at random* then we can be sure that all *known* and *unknown* factors or variables that might affect outcome are equally present across the groups, except by chance. Because any factor that is associated with outcome is the same in each group, the effects are balanced or cancelled out. Therefore, if we offer one of the groups an intervention condition that is different from the control or comparison condition the other group receives and we see a difference at outcome, we can be confident (if the sample size is large enough) that the difference is due to the intervention and alternative explanations can be ruled out. Although the basic concept of a RCT is straightforward, its

implementation is often rather more complex and requires careful thought in order to maintain the rigor of initial random allocation.

For the purposes of this chapter, we define *efficacy evaluation* as research that seeks to observe whether an intervention *can be effective* under optimum implementation conditions, and we define *effectiveness evaluation* as research that seeks to observe whether an intervention *is effective* when it is tested in authentic settings, where fidelity of implementation might be less than optimal. RCTs are not a panacea when researchers are addressing all such questions, but they are amenable to research questions that seek to investigate the relative efficacy or effectiveness of different policies or programs, such as different approaches for reducing criminal activity, different educational programs, different methods of improving voter turnout, and varying ecological or agricultural practices.

This chapter focuses on the use of RCTs in evaluating social policy interventions. In the field of social policy, public officials often intervene by introducing new policies, programs, and practices without their effectiveness having first been demonstrated using a rigorous design (such as an RCT). Biases and limitations that compromise the integrity of RCTs can be introduced at any stage in their design and conduct. However, all such biases and limitations also affect other research designs (quasi-experiments); and alternative methods usually have additional problems and sources of bias (Berk, 2005). The only exception to this is the regression discontinuity design. In this chapter, we highlight the key characteristics of well-designed RCTs and demonstrate how biases and limitations in the method can be avoided.

History of RCTs

The modern randomized controlled trial has its origins in the writings of Fisher in an agricultural context (Fisher, 1971). This research method was adopted for use with humans by education researchers in the 1930s (Lindquist, 1940; Walters, 1931), with medical researchers first using the technique about a decade later (Medical Research Council, 1944, 1948). Since those times, the use of RCTs in medical research has exploded, and now no new pharmaceutical product will be licensed until it has been tested in several large RCTs with human participants. This might be partly due to the many disasters that have occurred in the field of health care research when interventions, usually pharmaceutical interventions, were licensed in the absence of RCT evidence of their effectiveness in human participants (Silverman, 2004). Although occasionally using the RCT as their evaluative method, researchers

in other areas (such as education and criminal justice) have used RCTs much less frequently than health care researchers have. Mistakes in health care can often be counted in terms of morbidity levels; mistakes in other policy areas are less noticeable, but often as important. Society imposes regulations to ensure that a new cream for wart treatment must be tested in a large RCT, yet the same level of evidence is not required before new programs and practices are imposed on school or justice systems, even though an incorrect decision about education pedagogy or a criminal justice innovation may have a greater detrimental impact on society than the use of an ineffective wart treatment.

In recent times in the United States and the U.K., there has been a renewed interest in the use of experimental methods in education research—an interest driven by policymakers who are realizing that the RCT is the most robust method for providing firm answers to pressing questions of how to educate children and young people. Government policy initiatives are especially ripe for RCT evaluations. Changes to national or local policies incur huge costs and affect many thousands, if not millions, of citizens. Where possible, policy initiatives should be evaluated using the RCT design, as this could lead to more effective policies and large resource savings by avoiding the implementation of ineffective, but costly, policies.

Why Randomize?

It is possible for researchers to know that one intervention is more or less effective than an alternative intervention only when they can compare groups that are equal *in all known and unknown variables that relate to outcome* before each of these groups receives the experimental or the control intervention. Any differences observed thereafter can be safely assumed to be results of the intervention.

Randomization is simple. Researchers form groups of participants using a method akin to tossing a coin (although it is best to use a secure computer system to do the coin tossing). Random allocation is virtually the only method in which two or more groups that are similar in all known and unknown characteristics can be formed. If researchers form groups by virtually any other method, they are at risk of introducing selection bias. *Selection bias* occurs when the determination of group membership means that membership correlates with final outcome. For example, in a comparison of the performances of children attending private schools and children attending public schools, selection (particularly if it related to income) could bias or confound any results.

Trial Design

In its simplest form, the randomized controlled trial is used to assemble two groups through random allocation. One of the groups is then exposed to the treatment intervention, and the other (control or comparison) group is exposed to an alternative intervention or to no intervention. Both groups are then followed for a specified period of time and the outcomes of interest are observed. The nature of the control or comparison condition depends on the research question. For example, if the question is which of two methods for teaching reading is more effective, the comparison condition is one of the two experimental conditions. If the research question is whether a supplementary program for improving reading is effective, then the control condition will be a no-treatment condition.

Although, in theory, random allocation ensures comparable groups at baseline, bias can be introduced at any stage, including initial randomization. Even when two groups have been carefully formed through a rigorous randomization procedure, the groups might not remain comparable unless due diligence is observed in avoiding the introduction of all possible post-randomization biases. In the following sections, we highlight the main potential sources of bias (pre- and post-randomization) and illustrate how these can be minimized or avoided.

Biased Allocation and Secure Allocation

A crucial issue in the design of an RCT is ensuring that the initial allocation is truly random. There is a wealth of evidence, mostly from health care research, that some "randomized" trials are not RCTs at all (Berger, 2005). Anecdotal and statistical evidence has shown that some trials have suffered from subversion of the random allocation process (Berger, 2005; Hewitt and others, 2005; Schulz, Chalmers, Hayes, and Altman, 1995). In these cases, rather than using a random method of assigning participants to groups, the researchers used a systematic or selective approach to group allocation. This might have occurred through ignorance or might have been due to deliberate research misconduct. For instance, a survey of the extent of this problem in the health care field found examples where researchers had deliberately placed patients who were likely to fare better than others (because they were younger and fitter, for example) into the experimental group in order to "prove" that an experimental treatment was superior (Hewitt, Torgerson, and Berger, 2009). In addition, through ignorance, some researchers have overridden an initially

random allocation because, by chance, it had led to groups of different sizes, and they thought, in error, that they should correct this.

In the last ten to fifteen years, health care researchers have become aware that research misconduct does exist in the conduct of some randomized trials and that, consequently, rigorously designed trials should always use third-party randomization services to avoid this potential problem. Health care research is not likely to be the only field experiencing this problem; consequently, it should be accepted good practice to separate the randomization procedure from the researcher who is enrolling participants in the trial.

Contamination and Cluster Randomization

For many policy interventions, there is a huge potential for *contamination* of the control participants with the treatment condition. For example, in an experiment investigating the effectiveness of teacher praise in motivation, schoolchildren in the same classes were randomized as individuals to either receive or not receive praise from their teachers. However, the teachers could not prevent themselves from praising the children in the control group as well as the children in the treatment group (Craven, Marsh, Debus, and Jayasinghe, 2001). And one can readily envisage other examples where contamination might be a problem. For example, in one study, investigators wished to know whether an intervention to improve recycling household waste was effective (Cotterill, John, Liu, and Nomura, 2009). The intervention consisted of volunteers knocking on doors to urge householders to sort their waste for recycling. To evaluate this initiative, an RCT was undertaken. The outcome was the amount of sorted waste by household; however, randomizing by house might have introduced contamination, as neighboring households might (on seeing their neighbors recycling waste) have modified their own behavior. Consequently, street-level randomization, rather than house-level allocation, was undertaken.

When confronted with the likelihood of contamination, it is best to randomize at a level *higher* than the level of the individual. Classically, this has occurred in education research, where the class or school is usually the most appropriate unit of allocation (Lindquist, 1940). Such trials are called *cluster randomized* trials or experiments. Even when it is theoretically possible to randomize individuals, sometimes it is not feasible for practical or administrative reasons. For instance, in a trial evaluating the use of financial incentives to encourage attendance at adult literacy classes, cluster or class-based randomization was used (Brooks and others, 2008). This prevented control students from finding out about the incentives and the fact that they

were not going to receive them, as this knowledge might have altered their behavior: for example, they might have reduced their attendance if they felt resentful demoralization.

Ascertainment and Blinded Follow-Up

A major potential source of bias in RCTs is the way in which the assessment is conducted at follow-up. If the observer undertaking the posttest or follow-up measure is aware of individuals' group allocation, then he or she might consciously or unconsciously bias the outcome. It is important, therefore, that whenever possible, data collection be undertaken by an observer who is masked or blinded to the intervention status of the participants. Tests should be selected, administered, and assessed by personnel who are not aware of participants' group allocations. Failure to take this precaution can certainly introduce either conscious or unconscious bias on the part of those doing the data collection at follow-up.

Crossover and Intention to Treat

In many, if not most, randomized trials, some participants will receive a treatment different from the one to which they were assigned. In such cases it is tempting but scientifically incorrect to analyze participants by the condition that they *received* rather than that to which they were originally assigned. Randomization ensures that baseline confounders or covariates are equally distributed. If after randomization some participants in the control group gain access to the intervention, then these participants will invariably be different from those who do not gain access to the treatment condition. If the data from these participants are excluded from the analysis or, worse, included in the intervention group's data, this inclusion will result in bias. Therefore, their data should be included in the analysis *as if they had remained in their original randomized group.* This is known as *intention to treat* (ITT) analysis. Doing this will, of course, dilute any treatment effects toward the null. However, the worst that can happen in such an analysis is that it will be concluded that there is no difference when, in fact, there is a difference. If the data from the crossover participants were to be included with the intervention group's data, then it might be concluded that an intervention is beneficial when in truth it is harmful. Furthermore, if the crossover can be measured accurately, it will be possible to adjust the analysis using a statistical technique known as *complier average causal effect* (CACE) analysis (Hewitt, Torgerson, and Miles, 2006), which will under many circumstances provide an unbiased estimate of effect (as discussed later).

Attrition

Another potential threat to the reliability of an RCT is the problem of attrition. In many RCTs, some participants drop out of the study. If this dropping out is nonrandom, which it often is, then bias can be introduced. For example, if boys with low pretest scores are at increased risk of dropping out of an experimental group, then at analysis researchers might mistakenly conclude that the intervention either has no effect or makes things worse, as boys with low test scores are present in the control group, but their counterparts are no longer available in the experimental group. Consequently, it is important to ensure that attrition is kept to a minimum. Often participants confuse failure to comply with the intervention with dropping out of the study. Noncompliance is a problem, as noted previously; however, it is better to retain participants in the study for the posttests, even when they no longer comply with the intervention. When noncompliant participants are retained, bias due to attrition is avoided. Once it is explained to participants that failure to comply does not equate with dropping out and that their posttest data will be included in the analysis, they are usually happy to provide these data. When students change schools, it is worthwhile putting procedures into place to obtain their posttest outcomes by having research staff administer the posttest in their new schools. One quality assurance check for a good trial is that it has both low dropout rates overall and equivalent dropout rates between groups. If the dropout rates are different (for example, 20 percent in the control group and 10 percent in the intervention group), then there is a real worry that this might introduce attrition bias.

Resentful Demoralization: Preference Designs

Participants might have a preference for one of the options under evaluation, or policymakers might insist that their preferred intervention is rolled out to all. These issues pose problems for researchers evaluating a novel intervention using trial methodology. However, these problems are not insurmountable and, with a little thought, can often be addressed through careful trial design (Torgerson and Torgerson, 2008). Researchers do need to deal with the issue of treatment preference to avoid the potential for bias posed by this problem. They can do this through several different approaches. One approach is the *participant preference design*—also known as the Brewin-Bradley approach (Brewin and Bradley, 1989). In this design, participants are asked about their preferences before randomization, and only those who are indifferent to the conditions are randomized; the remainder receive their preferred

intervention and are then followed up in a nonrandomized study. (Note that in the field of education, issues of treatment preference may relate more to teachers or parents than to students.) In the Brewin-Bradley approach, those who are randomized provide an unbiased comparison, whereas the outcomes of the nonrandomized participants are likely to be biased.

Another alternative is the *randomized cohort design.* In this design, participants are initially recruited to a cohort study—that is, they consent to undertake pretests and to be followed up with posttests at regular intervals (Relton, Torgerson, O'Cathain, and Nichol, 2010). At recruitment they are also informed about potential interventions and asked whether they would consider using these at some point in the future. Participants who indicate an interest in one or more of the potential interventions are randomly assigned to produce a randomized experiment.

In some instances, it might be possible to randomize participants without their knowledge. For instance, an evaluation of a policy of reducing benefits to parents who do not have their children vaccinated before attending school randomized people, without their consent, to be either informed or not informed about this policy. An alternative to this is Zelen's method (Zelen, 1979), where participants are randomized and then *only those allocated to the novel intervention* are asked for their consent to receive the intervention. A major problem with these approaches is one of ethics: some ethics committees will refuse permission for a design where full informed consent is not part of the design. There are scientific problems with these methods as well. In the latter approach, if significant numbers of participants refuse the intervention and cross over to the control treatment, then study power is lost and, correspondingly, more participants are required. This occurs because the effects of the intervention will be diluted, and to maintain the randomized allocation, researchers will need to use ITT analysis (as noted earlier).

Waiting List and Stepped Wedge Designs

Alternative approaches to dealing with some of these issues include the *waiting list design* (where participants are informed that they will receive the intervention at a later date), and the *stepped wedge design.* Generally, researchers need a prima facie case for testing a new intervention in public policy, such as prior but not highly rigorous evidence of its effectiveness. A potential problem then faced by researchers undertaking an RCT is that the participants in the control group might be unhappy at being denied the promising intervention. In evaluating an intervention where the evidence is uncertain, there is no ethical problem; indeed, it is ethically correct to offer the participants a chance

(random possibility) to be enrolled in the most effective intervention, which might very well be the control condition. However, potential participants might not be convinced that the control intervention is as likely as the new intervention to be beneficial. When they anticipate a benefit from the experimental intervention, those allocated to the control group might suffer resentful demoralization and either refuse to continue with the experiment or deliberately or subconsciously do worse merely because they have been refused the new intervention. Thus, it might also be desirable to evaluate the "real-world" implementation (pragmatic trial) of an intervention that had previously been shown to be effective in a laboratory-type RCT (explanatory or efficacy trial).

In a waiting list study, participants are told explicitly that they will receive the intervention; however, some will receive it straightaway, and others will receive it later. It is then possible to evaluate the effectiveness of the intervention by measuring both groups at pretest, implementing the intervention in one group, and giving a posttest measurement, and after this giving the intervention to the participants in the control group. Consider an RCT undertaken by Brooks, Miles, Torgerson, and Torgerson (2006), for example. This study evaluated the use of a computer software package in a secondary school. Such packages were usually implemented arbitrarily, as there were insufficient laptop computers for all pupils to receive the software at the same time. For the evaluation, the researchers changed the arbitrary assignment to random allocation and adopted a waiting list design, which permitted a rigorous evaluation of this software package. Moreover, the use of the waiting list in this instance allowed all the children to receive the package eventually and might have reduced any demoralization, either among the children or among their teachers.

A special form of the waiting list design is known as the stepped wedge, or multiple baseline, method. Policymakers, particularly politicians, are often anxious to implement an intervention as soon as possible, which in some cases does lead to problems in evaluation. However, sometimes policymakers can be persuaded to adopt a staged approach in rolling out a program. It is then possible to randomize the order in which areas or units receive the intervention. More information on the stepped wedge design can be found in a recent systematic review of the method (Brown and Lilford, 2006).

The stepped wedge design differs from the waiting list design in that it operates as a series of waiting lists (Hayes and Moulton, 2009). Indeed, staged implementation might result in a more efficient method of rollout than the so-called big bang approach. For example, consider the implementation of a novel method of offender supervision by probation officers in the North

of England (Pearson, Torgerson, McDougall, and Bowles, 2010; Pearson, McDougall, Kanaan, Torgerson, and Bowles, 2014). Two areas wanted to implement the program, which required training probation officers in the new system. Researchers managed to persuade one area to implement the program office by office. Hence, the probation offices were randomly assigned to a waiting list so that implementation commenced in the first office, followed by a three-month gap, and then implementation began in the second office, and so on until all the offices had received the new training. In each three-month gap, the reoffence rates were monitored for all offenders attending all probation offices, thus enabling an unbiased estimate of the impact of the new service. In the second area, this approach was not taken and a big bang method was adopted. However, it was found that implementation was suboptimal because there were insufficient resources to deliver the training to all the sites in a short period of time. Process measures (such as measures of referrals to other services) indicated that the area that had adopted the stepped wedge approach was using services, such as alcohol counseling, more effectively than the big bang area was. Thus, in this case, adopting a rigorous method of evaluation ensured rigorous research *and* better training.

One potential weakness of the stepped wedge design is that it is necessary to measure outcomes at each step, that is, whenever an individual or cluster moves from the control to the intervention section of the wedge. This can be costly or intrusive to participants. Therefore, the stepped wedge design might work best when the outcomes are based on routinely collected data, such as national education assessments. Furthermore, it is important to monitor the implementation of the intervention in different sites in order to assess whether the nature of the intervention evolves with each step. This is important when interpreting the results, as the intervention in the last cluster could be substantially different from the intervention in the first cluster, as it inevitably changes with the increased experience of those implementing it.

Of course, a huge problem with any form of waiting list design is that, even if the evaluation shows that the intervention is ineffective, it might prove politically difficult to withdraw the intervention. Furthermore, considerable amounts of resources might have already been expended by giving the control group an ineffective intervention, which could have been avoided had a non-waiting-list design been used.

Design Issues in Cluster Randomized Trials

Most trials evaluating policy interventions use *cluster randomization.* As noted above, cluster randomization (for example, allocation of classes or schools

TABLE 7.1. EXAMPLE OF MINIMIZATION ON TWO IMPORTANT VARIABLES.

Variable	Intervention	Control
Large	3	2
Small	2	3
Rural	3	2
Urban	2	3

rather than individual students) will minimize contamination bias. However, there are some potential issues that need to be considered in the design of a cluster trial. First, multiple clusters are required in each group. A cluster trial consisting of two units (for example, two schools, two hospitals, two prisons) will not produce any valid results, as this is effectively the same as a two-person trial. Confounding at the level of the cluster (for example, special teacher characteristics or differences in offender populations) will not allow researchers to make any judgment about the effectiveness or otherwise of the intervention. Furthermore, statistical tests cannot be undertaken on a sample of two. Consequently, assignment of several clusters to each group is required. Some methodologists recommend at least seven clusters per group, whereas others state that five per group will suffice (Donner and Klar, 2000; Murray, 1998). However, greater numbers of clusters than these are usually required if researchers are to have the power to observe an important difference. Nevertheless, cluster trials tend to have fairly small numbers of units, usually fewer than fifty. In this instance, some form of restricted allocation method might improve the precision of the trial. One method of allocating small numbers of groups is *minimization* (Torgerson and Torgerson, 2007). In this method, a simple arithmetical algorithm is used to ensure that the clusters are allocated so that cluster-level covariates (for example, size or past performance) are balanced. Table 7.1 illustrates the key step in minimizing on two key variables. In the table, ten clusters have been randomized. The researchers' goal here is to ensure balance between the intervention and control groups on two key variables: unit size and whether the unit is in a rural or urban area. An eleventh cluster has the characteristics of being *large* and *rural*. To assess the group into which Cluster 11 should be allocated, researchers would add up the number of existing units across those two variables. As the table shows, this sums to six units for the intervention group and four units for the control group. To ensure that this imbalance is minimized, the eleventh cluster would be assigned to the control group. If the sums for the groups were exactly the same, then the allocation would be done

randomly. Minimization is particularly attractive when researchers wish to ensure balance across several variables. The use of minimization will ensure that statistical power is maximized and that the groups are comparable at baseline.

A more important issue that is frequently overlooked in the design of cluster trials is ensuring that the individual participants in each cluster are either a random sample or a census. Randomization of the clusters will ensure that selection bias is avoided if all, or a random sample, of the cluster members are included in the analysis. One way of introducing bias into a cluster trial is to randomize the clusters and then ask each cluster member to consent to taking part in the study. Inevitably there will be some individuals who do not consent, and if their choice is influenced by knowledge of the likely intervention, bias is likely to be introduced (Torgerson and Torgerson, 2008). To avoid this possibility, cluster members need to provide consent *before* randomization of the clusters occurs. In education research, this should be straightforward, as researchers could obtain parental and student consent prior to random allocation of the clusters.

Sample Size Issues

Many trials (particularly in the field of education) are relatively small (Torgerson, Torgerson, Birks, and Porthouse, 2005). Small trials are likely to have the consequence that researchers are able to observe a difference between the groups that might be of policy significance but is not of statistical significance. As a general rule, few education interventions, when compared against an *active* control (such as usual teaching or other business as usual), will yield an effect size bigger than half a standard deviation or half an effect size (that is, the difference in group posttest means divided by the standard deviation of the control group). Indeed, many effective education interventions might generate an effect size difference on the order of only a quarter or a fifth of a standard deviation. However, small effect sizes matter. Consider an effect size of 0.10 (a tenth of a standard deviation difference), which is considered small; however, if this effect occurred in a population taking an examination, it would lead to an additional 4 percent of that population achieving a passing grade. This could matter a great deal in a high-stakes testing situation.

The arithmetical calculation of a sample size is relatively straightforward, and most statistical packages have a sample size function. Indeed, several computer packages that can be downloaded for free work quite satisfactorily (for example, PS Power from the Biostatistics Department at Vanderbilt University: http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/PowerSampleSize).

Determining what difference is worthwhile to policymakers and consumers requires a complex judgment. Such a difference would be influenced by the cost of the intervention, its ease of implementation, and various political and social factors. For instance, the aforementioned recycling program was relatively costly compared with simply sending leaflets to all households and would, all other things being equal, have needed to produce a relatively large effect size in order to justify its additional cost.

If a cluster design is being proposed, then the relationship between members of the cluster needs to be taken into account in the sample size calculation. This will generally lead to an increase in the number of individuals in the trial on the order of at least 50 percent, if not more. To adjust for the effects of clustering, the standard sample size needs to be multiplied by $[(m-1)\ 3\ ICC]\ 1\ 1$, where m is the size of each cluster and ICC is the intracluster correlation coefficient, which can be obtained from previously published studies. As a rule of thumb, ICC values are generally between 0.01 and 0.02 for human studies (Killip, Mahfoud, and Pearce, 2004). A final point concerning cluster studies is that there is a diminishing marginal return from increasing cluster size beyond twenty to thirty individuals per cluster (see the graph in Brown and others, 2008; also see Campbell and others, 2004).

Increased Power for Very Little Cost

It is almost automatic when randomizing participants or clusters to different treatments in a trial to attempt to have the same number of cases in each treatment group—a 1:1 allocation ratio of intervention group to control group. This tradition has grown up because a 1:1 ratio, for any given sample size, usually ensures maximum statistical power—where *power* is the likelihood of correctly finding a difference between the groups for a predetermined effect size. However, where resource shortages limit the number of people who can be offered the intervention treatment, power can be increased by randomly allocating more participants to the control group, thereby increasing the total sample size. For example, there might be sufficient resources to offer an intervention to only sixty-three participants. If equal allocation were used, then the sample size would be constrained to 126 participants. This would give 80 percent power to detect a difference of half a standard deviation between the two groups. However, if the allocation ratio is set to 2:1, then 189 participants can be randomized, with 126 in the control group and sixty-three in the intervention group, which will increase power to 90 percent. If this is done, statistical power is increased at little or no additional cost when the control group participants simply receive normal treatment anyway. The allocation ratio can be set as high as is desirable, although once it exceeds 3:1 the extra increase in

power tends to be slight and it might not be worth the effort of following up a much larger sample.

In summary, increasing the size of the control group in this way can give increased power for very little cost. Of course, if the *total* sample size is constrained, then using unequal allocation will reduce power—although not by much unless the ratio exceeds 2:1. For example, in a trial with 100 participants, if thirty-two were allocated to one group and sixty-eight allocated to the other group, the decline in power would be only 5 percent, compared with the power in a situation where fifty were allocated to each group. This loss of power might be worthwhile if considerable resource savings can be achieved.

Analytical Issues

The statistical analysis of most randomized controlled trials is relatively simple. Because selection bias has been minimized through the initial randomization, there is, in principle, no reason to use complex multivariate methods. However, in some trials, particularly in education, the pretest has a very strong relationship with the posttest. Similarly, in the aforementioned recycling trial, previous recycling behavior correlated strongly with future behavior. Therefore, it is desirable to control for pretest values in such trials. This is particularly important in order to improve statistical power. This relationship can be used either to increase power for a given sample size or in order to use smaller sample sizes. For example, a pretest-posttest correlation of 0.7 (that is, an R^2 value of 0.49) leads to an approximate halving of the required sample size (for any given power, significance value, and effect size). Often correlations are in excess of 0.7 in education trials, and this will further drive down the required sample size.

If cluster trials are undertaken—if randomization is, for example, at the class, school, hospital, prison, or district level—then the clustering needs to be taken into account in any analysis. The simplest way of doing this is to calculate the group means from the cluster means and perform an unpaired t test comparing the group means. To take covariates into account, such as pretest scores, one could extend this analytical framework to use a regression approach, taking into account cluster-level pretest scores. Note that a t test or simple regression analysis of individuals who have been randomized in clusters is always incorrect, no matter how few or how many clusters have been randomized. There are alternative approaches, including multilevel modeling, that also adjust for clustering and have some advantages, particularly when there is a complex design with many different levels (for example, children within classrooms, within schools, or within areas).

Generalizability or External Validity

In this chapter, we have focused on the internal validity of the RCT. That is, we have looked at whether researchers and stakeholders can rely on the results of a trial within an experimental sample. This is the correct focus, because without this internal reliability the results cannot be applied outside the study sample. Yet, it is important that the results of a trial should be transferable. The main point of an RCT is to influence policy beyond any particular evaluation. In this section, we look at some of the issues that need to be considered in order to make trials externally valid.

One of the criticisms of RCTs is that they are often not conducted in real-world, authentic settings. Many RCTs of education interventions, for example, are conducted in the laboratory setting of a university psychology department. However, trials are needed that have, first, been conducted with student participants in educational settings that are representative of normal educational practice and, second, replicated in diverse educational settings, as this will increase their generalizability. Consequently, pragmatic trials are needed where whole classes or schools are allocated to either receive a new program or to continue with business as usual. These schools need to be chosen to ensure that they represent the general population of schools. Indeed, this is where social science trials are different from medical experiments. Health care trials—specifically pharmaceutical trials—are more likely to transfer beyond their experimental population than education trials are. For example, an educational program developed in the United States might not apply to U.K. students and, even if we ignore language differences, is unlikely to apply further afield. The reason for this is that educational achievement can be profoundly affected by cultural and socioeconomic factors. This also applies to other fields. Criminal activity and types of offending behavior vary significantly between countries: for example, violent crimes involving the use of firearms are more prevalent in the United States, whereas those involving knives are more prevalent in the United Kingdom. Thus, interventions to prevent violent crime might require a focus in the United States that is different from the focus chosen in the United Kingdom.

Quality of Randomized Trials

Earlier we discussed some of the potential problems with undertaking RCTs. Many RCTs do not describe their methods in sufficient detail for outsiders to be sure that, for example, randomization was undertaken in a robust manner. Similarly, significant numbers of RCTs do not use an independent, blinded

follow-up or ITT analysis (Torgerson, Torgerson, Birks, and Porthouse, 2005). We have proposed elsewhere that when reporting their RCTs, researchers should adopt a modified version of the CONSORT statement used in health care research (Torgerson and Torgerson, 2008). The CONSORT statement is a checklist of twenty-two items that correspond to quality issues in the design, conduct, and reporting of RCTs and that need to be addressed in the trial manuscript if the reader is to be assured of the RCT's internal and external validity. Items include justification of the sample size, method of randomization, description of the population and intervention, and use of confidence intervals. Various papers about the CONSORT statement can be accessed online (www.consort-statement.org). Health care research methodologists developed the statement in response to the substantial number of poorly reported RCTs in health care research. Subsequently, it has been adopted by all the major health care journals, the major psychology journals, and more recently, some education journals, with other researchers such as political scientists also starting to use it (Cotterill, John, Liu, and Nomura, 2009).

Barriers to the Wider Use of RCTs

Many arguments are made against the wider use of RCTs and experimental methods when evaluating education interventions, some of which do not withstand sustained scrutiny. Policy implementation without evaluation using randomized designs is often justified on the ethical grounds that it is unethical to withhold promising interventions. Policies introduced by one set of politicians can easily be undone by a future group, particularly if there is no rigorous evidence to support their continuance. However, it is unethical to widely implement a policy that increases costs and might not result in benefit. For example, an enhanced sex education program in Scotland was implemented in state schools before the results of the trial on unwanted pregnancies became known. The trial showed that the program led to an *increase* in unwanted pregnancies and cost fifty times more than the existing program (Henderson and others, 2007).

Cost is often cited as a reason for not undertaking RCTs. Yet the cost of not doing them is in the long run likely to be greater. Furthermore, a carefully planned RCT is not necessarily expensive, and the value of an RCT can be estimated using value of information methods (Claxton and others, 2004).Involvement with stakeholders at an early stage is crucial. One of the reasons that quasi-experiments are so widely used is that the policy has already been implemented and then researchers have been asked to evaluate the decision after rollout. When early engagement with stakeholders occurs, then

sometimes randomization can be implemented. For example, the involvement of researchers early on *before* a novel policy in the probation service (described earlier; Pearson, Torgerson, McDougall, and Bowles, 2013; Pearson, McDougall, Kanaan, Torgerson, and Bowles, 2014) was implemented allowed them to persuade one of the probation services to use random allocation in its implementation strategy. Consequently, the early engagement of research and the appropriate stakeholders allowed a rigorous implementation and evaluation in one of the two districts. Note, however, despite early engagement between researchers and stakeholders, one of the districts refused to implement random allocation. Nevertheless, a 50 percent success was better than nothing!

Conclusion

The widespread use of random allocation is one of the most important methodological contributions to health and social science research of the last century. In education and additional fields other than health care, there is increasing interest in using the approach more widely. Although this renewed interest is welcome, it is necessary to ensure that trials are conducted to the highest standard; otherwise, there is a risk that their integrity will be compromised. Health care trials can contribute to informing our methodological deliberations. Although the traditional, placebo-controlled drug trial is unlike most educational program evaluations, there are many nonpharmaceutical health care research interventions that are similar to educational programs in key ways: for instance, health promotion or health education programs. For example, health care researchers have undertaken large, cluster randomized trials involving schools to assess new health promotion programs. Evaluations of literacy or numeracy programs can use the same method. If researchers in the field of health promotion can design and conduct rigorous trials of new programs in their field, there should be no methodological barrier preventing researchers from doing the same in the field of education. Many of the methodological advances in health care trial methods can be applied to educational program trials, and there is now published guidance for the evaluation of complex interventions in health care (Medical Research Council, 2009). One example is the need to monitor the fidelity with which an intervention is implemented.

It is important that trials that are undertaken be rigorous; otherwise funders of randomized controlled trials might turn away from them in the future.

Consequently, at the same time that we urge evaluators to use the RCT, we note that it is equally important that they use the most rigorous methods available.

References

Berger, V. W. *Selection Bias and Covariate Imbalances in Randomized Clinical Trials.* Hoboken, NJ: Wiley, 2005.

Berk, R. A. "Randomized Experiments as the Bronze Standard." *Journal of Experimental Criminology,* 2005, *1,* 417–433.

Brewin, C. R., and Bradley, C. "Patient Preferences and Randomised Clinical Trials." *British Medical Journal,* 1989, *299,* 313–315.

Brooks, G., Burton, M., Coles, P., Miles, J., Torgerson, C., and Torgerson, D. "Randomized Controlled Trial of Incentives to Improve Attendance at Adult Literacy Classes." *Oxford Review of Education,* 2008, *34,* 493–504.

Brooks, G., Miles, J.N.V., Torgerson, C. J., and Torgerson, D. J. "Is an Intervention Using Computer Software Effective in Literacy Learning? A Randomised Controlled Trial." *Educational Studies,* 2006, *32,* 133–143.

Brown, C. A., and Lilford, R. J. "The Stepped Wedge Trial Design: A Systematic Review." *BMC Medical Research Methodology,* 2006, *6,* 54.

Brown, C. A., and others. "An Epistemology of Patient Safety Research: A Framework for Study Design and Interpretation. Part 2: Study Design." *Quality & Safety in Health Care,* 2008, *17,* 163–169.

Campbell, M. K., and others. "Sample Size Calculator for Cluster Randomised Trials." *Computers in Biology and Medicine,* 2004, *34,* 113–125.

Claxton, K., and others. "A Pilot Study on the Use of Decision Theory and Value of Information Analysis as Part of the NHS Health Technology Assessment Programme." *Health Technology Assessment,* 2004, *8*(31), 1–103.

Cotterill, S., John, P., Liu, H., and Nomura, H. "Mobilizing Citizen Effort to Enhance Environmental Outcomes: A Randomized Controlled Trial of a Door-to-Door Recycling Campaign." *Journal of Environmental Management,* 2009, *91,* 403–410.

Craven, R. G., Marsh, H. W., Debus, R. L., and Jayasinghe, U. "Diffusion Effects: Control Group Contamination Threats to the Validity of Teacher-Administered Interventions." *Journal of Educational Psychology,* 2001, *93,* 639–645.

Donner, A., and Klar, N. *Design and Analysis of Cluster Randomization Trials in Health Research.* London: Hodder Arnold, 2000.

Fisher, R. A. *The Design of Experiments.* New York: Hafner, 1971.

Hayes, R. J., and Moulton, L. H. *Cluster Randomised Trials.* London: Chapman & Hall, 2009.

Henderson, M., and others. "Impact of a Theoretically Based Sex Education Program (SHARE) Delivered by Teachers on NHS Registered Conceptions and Terminations: Final Results of Cluster Randomised Trial." *British Medical Journal,* 2007, *334,* 133.

Hewitt, C. E., Torgerson, D. J., and Berger, V. "Potential for Technical Errors and Subverted Allocation Can Be Reduced If Certain Guidelines Are Followed: Examples from a Web-Based Survey." *Journal of Clinical Epidemiology,* 2009, *62,* 261–269.

Hewitt, C. J., Torgerson, D. J., and Miles, J.N.V. "Taking Account of Non-Compliance in Randomised Trials." *Canadian Medical Association Journal,* 2006, *175,* 347–348.

Hewitt, C., and others. "Adequacy and Reporting of Allocation Concealment: Review of Recent Trials Published in Four General Medical Journals." *British Medical Journal*, 2005, *330*, 1057–1058.

Jacob, B. A., and Lefgren, L. "Remedial Education and Student Achievement: A Regression-Discontinuity Analysis." *Review of Economics and Statistics*, 2004, *86*(1), 226–244.

Killip, S., Mahfoud, Z., and Pearce, K. "What Is an Intracluster Correlation Coefficient? Crucial Concepts for Primary Care Researchers." *Annals of Family Medicine*, 2004, *2*(3), 204–208.

Lindquist, E. F. *Statistical Analysis in Educational Research.* Boston, MA: Houghton Mifflin, 1940.

Medical Research Council. "Clinical Trial of Patulin in the Common Cold." *Lancet*, 1944, *2*, 370–372.

Medical Research Council. "Streptomycin Treatment of Pulmonary Tuberculosis: A Medical Research Council Investigation." *British Medical Journal*, 1948, *2*, 769–782.

Medical Research Council. *Developing and Evaluating Complex Interventions: New Guidance.* London: Medical Research Council, 2009.

Murray, D. M. *Design and Analysis of Group-Randomized Trials.* New York: Oxford University Press, 1998.

Pearson, D.A.S., Torgerson, D. J., McDougall, C., and Bowles, R. "Parable of Two Agencies, One Which Randomizes." *Annals of the American Academy of Political and Social Sciences*, 2010, *628*, 11–29.

Pearson, D.A.S., McDougall, C., Kanaan, M., Torgerson, D. J., and Bowles, R. "Evaluation of the Citizenship Evidence-Based Probation Supervision Program Using a Stepped Wedge Cluster Randomized Controlled Trial." *Crime and Delinquency*, 2014,

Relton, C., Torgerson, D. J., O'Cathain, A., and Nichol, J. "Rethinking Pragmatic Randomized Controlled Trials: Introducing the 'Cohort Multiple Randomized Controlled Trial' Design." *British Medical Journal*, 2010, *340*, c1066.

Schulz, K. F., Chalmers, I., Hayes, R. J., and Altman, D. G. "Empirical Evidence of Bias: Dimensions of Methodological Quality Associated with Estimates of Treatment Effects in Controlled Trials." *JAMA*, 1995, *273*, 408–412.

Shadish, W. R., Cook, T. D., and Campbell, T. D. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Boston, MA: Houghton Mifflin, 2002.

Silverman, W. "Personal Reflections on Lessons Learned from Randomized Trials Involving Newborn Infants from 1951 to 1967." *Clinical Trials*, 2004, *1*, 179–184.

Torgerson, C. J., and Torgerson, D. J. "The Use of Minimization to Form Comparison Groups in Educational Research." *Educational Studies*, 2007, *33*, 333–337.

Torgerson, C. J., Torgerson, D. J., Birks, Y. F., and Porthouse, J. "A Comparison of Randomised Controlled Trials in Health and Education." *British Educational Research Journal*, 2005, *31*, 761–785.

Torgerson, D. J., and Torgerson, C. J. *Designing Randomised Trials in Health, Education and the Social Sciences.* Basingstoke, UK: Palgrave Macmillan, 2008.

Walters, J. E. "Seniors as Counselors." *Journal of Higher Education*, 1931, *2*, 446–448.

Zelen, M. "A New Design for Randomized Clinical Trials." *New England Journal of Medicine*, 1979, *300*, 1242–1245.