CHAPTER 21

Randomized Block Designs

BARAK ARIEL AND DAVID P. FARRINGTON

INTRODUCTION

In randomized controlled trials, the experimental and control groups should be relatively balanced. This balance is achieved through a process of random assignment of study participants into treatment and control groups because the researcher can control for any preexisting differences between the two groups. Because participants are assigned to groups by chance, the likelihood that the experimental group will turn out to be different from the control group is significantly minimized when compared with a nonrandom assignment procedure.

But while this simple random assignment procedure is likely to provide an overall balance, it does not guarantee that there will be complete balance on any particular trait (see Altman et al. 2001; Berger and Exner 1999; Proschan 1994; Senn 1989, 1994; Wei and Zhang 2001; Weisburd and Taxman 2000). Unbalance is usually associated with a lower statistical power of the research design (see Lipsey 1990; Weisburd and Taxman 2000). So, interpreting the results from an unbalanced trial may lead to reaching biased conclusions about the true outcome effect of the tested intervention (Torgerson and Torgerson 2003). When the researcher anticipates that the groups may be unbalanced, certain measures should be taken.

It was (and perhaps still is) commonly assumed that increasing the size of the sample is associated with more balance and, in turn, with higher statistical power. But, as Weisburd et al. (1993) have shown, there is in practice little relationship between sample size and statistical power. Larger samples are likely to include a wider diversity of participants than smaller investigations. Because it is necessary to establish very broad eligibility requirements in order to gain a larger number of cases, large trials attract a more diverse pool of participants. This increases the variability in the data as there is more "noise," which makes it difficult to detect the effect of the treatment. Therefore, the design benefits of larger trials may be offset by the implementation and management difficulties they present.

Clinicians and statisticians have developed different research designs to more adequately handle unbalanced groups or high variability, such as the randomized block design (RBD). Generally speaking, if we know before we administer the treatment that certain participants may vary in significant ways, we can utilize this information to our advantage. We should use a statistical design in which participants are not simply randomly allocated directly into experimental and control groups. Instead, they are randomly allocated into groups within blocks, in a design commonly referred to as the RBD.

A.R. Piquero and D. Weisburd (eds.), *Handbook of Quantitative Criminology*,
DOI 10.1007/978-0-387-77650-7_21,

Though generally ignored in criminal justice, variations of this design were implemented in some of the most influential studies in recent years, such as the Minneapolis Hot-Spots Experiment (Sherman and Weisburd 1995), the Jersey City Drug Market Experiment (Weisburd and Green 1995), the Jersey City Problem-Oriented Policing Experiment at violent places (Braga et al. 1999) and others. Recently, Weisburd et al. (2008) have used this design in an experiment on the effects of Risk-Focused Policing at Places (RFPP) approach to preventing and reducing juvenile delinquency. Before administering the treatment, they found a considerable variability in the characteristics of violent places selected for the study. This variability jeopardized the statistical power of the evaluation. For example, if we assume that police are more likely to be successful in addressing problems in the school domain than they are in the community domain, then by using simple randomization, the experimental group comprised of places in which this approach is implemented may result in more places with high risk factor scores in the school domain by chance alone. In this case, observation of treatment impacts will be more difficult to detect when there is a large variability or "noise" in the assessment of study outcomes. The data were thus more adequately managed under a RBD.

In the Jersey City Drug Market Experiment, Weisburd and Green (1995) tested the effects of a drug enforcement strategy developed for the drug market in Jersey City. The researchers identified 56 places of increased drug activity, which were then randomized in statistical blocks to experimental and control conditions. This, as well as other hot-spot experiments (e.g., Sherman et al. 1989; Sherman and Weisburd 1995), benefit from a RBD because it decreases the variations that characterize places, such as in terms of the drug activity, structural and cultural characteristics in the identified hot spots. An examination of the distribution of arrest and call activity in the hot-spots before the experiment revealed that the places sample falls into four distinct groups: very high arrest and call activity, and the high, medium, and low activity. Therefore, hot spots were randomly allocated to experimental and control conditions within each of these four groups.

RBDs can also be used in non place-based experiments. For example, similar issues arise in analyses of individuals in schools, prisons and other environments, that can also be regarded as blocks, for the purpose of the analyses (such as in Farrington and Ttofi, 2009). A large randomized controlled trial was conducted in order to test the hypothesis that differences in wording of letters sent to taxpayers in Israel would affect various aspects of their taxpaying behavior, such as how much money they are willing to report and pay to the tax authority (Ariel 2008). Nearly 17,000 taxpayers were randomly assigned to different groups, each receiving a different type of letter. However, a large sample of taxpayers introduced high variability. For example, the sample included very poor taxpayers as well as very rich taxpayers. The size of the sample also increased the likelihood that in one of the study groups, there would be a disproportionally larger number of extremely rich taxpayers, so their effect on the results could skew the conclusion. Therefore, the sample was divided, before random assignment, into blocks of income levels. Within each of these income-level blocks, the participants were then randomly allocated into the different letter groups. Therefore, this procedure allows for measurement of not only the overall effect of letters, but also of specific effects within the blocks (i.e., effect of letters on participants with different income levels) the data can then be analyzed using an ordinary analysis of variance and other commonly used statistics. However, over the years more complex methods have been suggested for analyzing these designs (see Chow and Liu 2004; Dean and Voss 1999; Friedman et al. 1985; Lachin 1988b; Matts and Lachin 1988; Schulz and Grimes 2002; Yusuf et al. 1984), but these go beyond the scope of this chapter.

In this chapter, we will explore the RBD. We will briefly review some the threats that cannot be adequately resolved when using the simple randomization procedure, but can be when using RBD. We will then show how the design works, and introduce four particular models of this design. Each type is more adequate for certain test conditions. We will conclude by discussing more general issues related to the design, particularly why the benefits of using it are too large to ignore.

LIMITATIONS OF SIMPLE RANDOMIZATION DESIGNS

The simple, or complete, randomization design (CRD) is the most prevalent method of random assignment in criminal justice research (Ariel 2009). Under CRD, a randomly chosen subset of units n_a out of n units is assigned to treatment a and $n_b = n - n_a$ units are assigned to treatment b. In this way, the experimental and control groups should be equivalent, in all measured and unmeasured extraneous variables.

But as we have reviewed in the introduction, CRDs are less adequate to deal with certain statistical threats, such as high variability between the participants in the study, or when the researcher anticipates that the groups will be unbalanced. When the data are indeed characterized by high variability or imbalance, it decreases the ability of the researcher to determine the true effects of the intervention. There is a large body of literature which discusses the reasons for, as well as the adverse implications of, this threat (Kernan et al. 1990: 20, Lachin 1988a, b; Palta and Amini 1982; Lachin et al. 1988; Pocock 1979). Some of these studies further show that CRDs are less adequate to deal with certain practical constraints, such as those associated with the studies in which eligible participants are assigned to either treatment(s) or no-treatment in a sequential order (Pocock and Simon 1975; Proschan 1994). In light of these and other limitations of the simple randomization technique, alternative random allocation procedures have emerged.

The Randomized Complete-Block Design

The Randomized Complete-Block Design (RCBD), sometimes referred to as the simple complete-block design, is a frequently used experimental design in biomedical research (Cochran and Cox 1957; Lagakos and Pocock 1984; Abou-El-Fotouh 1976; see also Hill 1951; Fisher 1935; Ostle and Malone 2000: 372), but it is quite rare in criminal justice research (Ariel 2009). It is often adequate when there are "several hundred participants" or less (Friedman et al. 1985: 75). The experiments reviewed in the introduction are all examples of this design.

The RCBD is used in order to decrease the variance in the data (Lachin 1988a). Unlike CRDs, where units are unrestrictedly distributed at random to either treatment or control (or more than two groups, as the case may be), under the RCBD model, units are allocated randomly to either treatment or control within pre-identified blocks. The blocking process is established based on a certain criterion which is intended to divide the sample, prior to assignment, into subgroups that are intended to be homogeneous (Hallstrom and Davis 1988; Simon 1979). Then, within blocks, units are randomly assigned to either treatment or control conditions.

It is important that the administration of the treatment in terms of potency, consistency and procedures, is identical in each block. This means that the overall experimental design is replicated as many times as there are blocks, and each block can be viewed as a disparate yet identical trial within the overall study (Rosenberger and Lachin 2002). If the administration of the treatment is different in each block, it becomes rather difficult to analyze the treatment effect (see Ostle and Malone 2000; Matts and Lachin 1988; Rosenberger and Lachin 2002).

THE HYPOTHETICAL CASE OF A RCBD: THE DRUG TREATMENT EXPERIMENT. Consider the following hypothetical trial, as a way to show the benefits of blocking. Imagine a trial with one experimental group consisting of 52 drug-addicted offenders, treated in an antidrug program. Another group of 48 drug-addicted offenders serve as the control group, receiving no drug treatment. The allocation to treatment and control is done randomly. The experiment is conducted in order to evaluate the merits of the program, where success is nominally defined by a decrease in drug use (on a scale of 1–8, 8 being the highest and 1 being lowest). However, unlike other drug treatment programs, this particular treatment is very costly. Therefore, unless very high success rates are registered, the facilitators will be unlikely to recommend its implementation in the future. All eligible participants selected for this trial are known to be drug abusers before entering the program. Prior to the treatment, the drug use level was not statistically different between the two groups (both averaging about 5.85 on the said scale). At the 6-month follow-up period, drug use was measured again for both groups. Results show that the program was successful: there was an overall 30% reduction in the treated experimental group, when compared with the untreated control group. A visual depiction is presented in Fig. 21.1. At the same time, considering the costs of the program, the policy implication is to not recommend further implementation of the program.

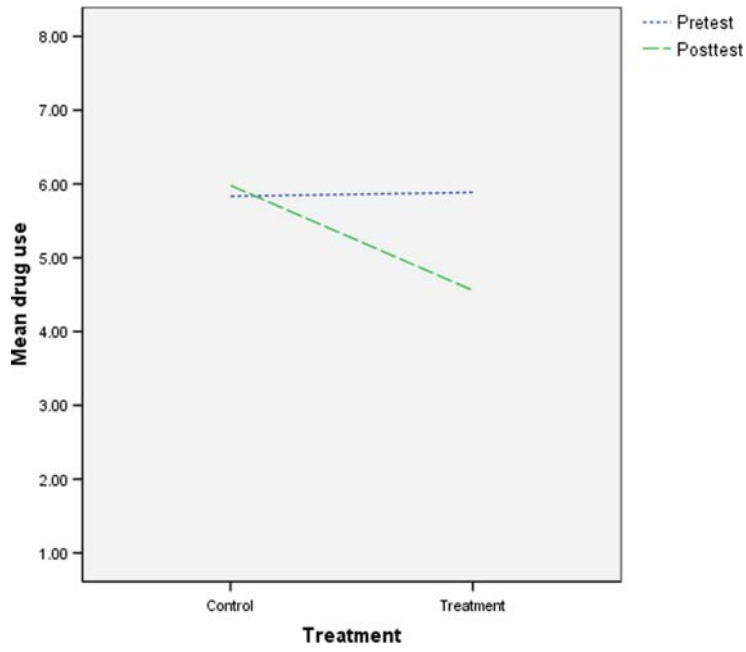

FIGURE 21.1. Hypothetical drug treatment trial: Phase I – CRD.

However, a closer look at the distribution of the pre-test scores indicates that the sample is comprised of drug abusers of four different types of drugs (with no crossovers): marijuana, MDMA, alcohol, and heroin. Generally speaking, it could be hypothesized that drug abusers of these different drugs are not addicted in the same way (based on both pharmacological and psychobiological qualities of these substances). Therefore, the prospective success rate of the drug intervention program is also hypothesized to be disparate for each subgroup; it is likely that getting clean from a heroin addiction or alcohol addiction is more difficult than from marijuana or MDMA. Thus, dividing the overall sample based on the type of drug should produce more homogeneous subgroups, whose chance to get clean is inevitably different across these subgroups. Even if the strength of the effect within each block is similar, blocking should give better estimates of the treatment effect, because of the decreased intrablock variance, or noise. Thus, a second experiment is conducted.

In the second experiment, the treatment is delivered in the very same intensity and method it was delivered when a complete randomization design was used. Nor do we make any modifications to the sampling of participants. Assuming that a similar group of participants was randomly selected, the overall 30% difference is also expected to be registered. However, because the treatment estimate is also comprised of the variability, the effect is larger within each block. In other words, the intrablock variance of the prognostic outcome is now less than the variability across the blocks, which makes it easier to detect stronger treatment effects. This is depicted in Fig. 21.2.

As can be seen, the treatment-to-control magnitude of difference is the same across blocks, but the intrablock difference is different, because there is less variance in the data. Furthermore, this disaggregation of the data into blocks indicates that the Heroin Block is less receptive to the treatment (6% decrease in drug use), as arguably expected from this

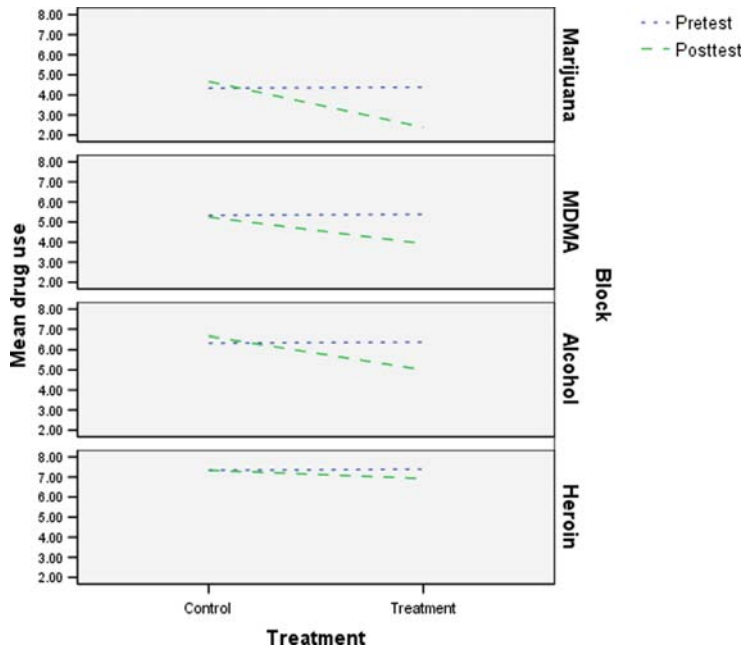

FIGURE 21.2. Hypothetical drug treatment trial: Phase II – CRBD.

subgroup, compared for example to the Marijuana Block or the MDMA Block (51% and 34% decrease, respectively). Therefore, the policy implication could then be to continue using the program for certain addictions in light of the high success rate, but not for others with a lower success rate.

This hypothetical trial was meant to show the conceptual groundwork for the RCBD. It shows how the researcher can benefit from blocking the data before random assignment, in a way that does not alter in any way the intervention or the actual allocation of participants to treatment or nontreatment. By blocking the participants, the potency of the drug program did not change. However, because the data are now classified in a way that homogenizes participants based on a qualitative criterion (type of drug), the variance within each block is smaller than the variance across the overall study.

Analyzing Randomized Block Designs

ORDINARY RBD. There are different models of RBDs. In each one, there is a particular kind of statistical analysis that is most adequate. In this section, we introduce one that is commonly used in medical research: the ordinary RBD model. Here, we analyze the data by looking at effects caused by the treatment factor, by the blocking factor, or by the interaction between the treatment factor and the blocking factor. This design is particularly adequate when we believe that there is some sort of association between the trait which is used as a criterion for the blocks (e.g., crime level) and the trait(s) of the treatment condition (e.g., hotspot policing). There are cases when this assumption of association between the treatment and the blocking is not necessary, and therefore is ignored in the statistical analysis. But these scenarios are arguably more complex than the scope of this chapter, so specialized references should be consulted (see Canavos and Koutrouvelis 2008; Gacula 2005; Hoshmand 2006: 14; Hinkelmann and Kempthrone 1994; Kepner and Wackerly 1996; Liebetrau 1983; Matts and Lachin 1988; Milliken and Johnson 1996; Mottonen et al. 2003; Ostle and Malone 2000; Rosenberger and Lachin 2002).

One common method to analyze the RBD is by using an ordinary two-way ANOVA, or through a general linear model which will include an interaction term between the fixed treatment factor(s) and a random blocking factor term. The model can therefore be described as follows:

$$Y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \varepsilon_{ijk}$$

In this equation, Y_{ijk} represents any observed value for treatment factor (i), the blocking factor (j), and any additional effect that may exist because of the interaction between these two factors (k). The model can be, at least in theory, extended to account for an endless number of treatments and blocks. However, there is a limited number of interventions one research project can study, and naturally the number of blocks is directly affected by the number of variables on which the data can be blocked. In practice, however, it is usually the case that only one blocking factor will be used and it is often recommended that the number of blocks should not exceed the number of treatments (Gill 1984; Matts and Lachin 1988). There are also those who recommend to use only one treatment and one control conditions in each block, because the test statistic is considered more "stable" with two groups (see Canavos and Koutrouvelis 2008; Gacula 2005; Ostle and Malone 2000). Most importantly, however, is that each block will contain all treatments and that each treatment occurs an equal number of times in each block – hence the "completeness" characteristic of this design (Chow and Liu 2004: 136–140). Many of the experiments on policing crime in places are constructed this way, such as the ones reviewed in the introduction.

This notwithstanding, there are cases when more than one treatment factor or blocking factor should be used when there is good theoretical basis for further division of the sample. In a study of defiance, for example, it would be reasonable to block the data on several blocking factors, such as (a) categories of criminal background, (b) categories of psychological variables, and also (c) the amount of legitimacy the sanctioned offender ascribes to the sanctioning agent's behavior (Sherman 1993). But, how many blocking factors should be allowed before the design becomes too "messy" (see Milliken and Johnson 1996: 259)? Friedman et al. (1985: 69) claim that, for studies of 100 participants, blocking using up to 3 factors is still considered manageable. Pocock and Simon (1975) make a similar argument, showing that beyond three factors, imbalances can result in incomplete filling of blocks, which complicates the analysis.

Going back to the equation, μ signifies a grand mean, derived by the values of all units in the study and is considered a constant. The effect of the treatment is marked by τ_i, and the β_j element marks the blocking variable effect. The possible interaction between these two is shown as $(\tau\beta)_{ij}$.

Lastly, the error is represented by the ε_{ijk} term, which includes the two sources of residual variance that were not explained by the model. First, there is variance related to random and normally distributed differences in the sample or the population from which the sample was selected. These differences are assumed to equal out when the data are aggregated. Second, there is variance the researcher did not identify, but that has a systematic impact on the relationship between the treatment and the effect. The consequence of these errors is loss of statistical power of the test. So, researchers should try to avoid them as much as possible, for example by anticipating them in advance and creating better or more blocks (see Dean and Voss 1999: 299–301).

Randomized Incomplete Block Designs. The RCBDs described earlier are characterized by "completion," in the sense that each block is complete, including all treatments and all control slots. However, there are instances when it may not be possible to apply all treatments in every block. This is particularly the case when there are many treatments tested in one trial, or with many replications. On the one hand, a larger number of replications is usually a good thing because it translates into greater precision of the treatment effect estimate. However, this is not always feasible. For example, in a trial testing the effect of various prison-based treatments for sexual offenders, because of their level of perverted sexual cognition, some offenders may not be capable of entering the various programs offered, after the stage of random assignment. Another example may be a trial on the effect of intensive police patrols on crime reduction in a certain metropolitan city. The police may be unable to allocate enough police patrol units to the intensified crime hot-spots because of some other major police activity (such as patrolling in the Olympics). These examples can be better handled through a design that takes into account this incompleteness.

If all possible comparisons between different treatments are seen as important for the researcher, then each pair of treatments must appear together in a block the same number of times, and each treatment must be observed the same number of times in a block. Otherwise, it becomes difficult to analyze the effect of each treatment.

One way to somewhat circumvent this problem is by implementing an intent-to-treat (ITT) approach, which works by "ignoring empty cells," under a relatively relaxed assumption that missing data are proportionally distributed across study groups or blocks. However, ITT would usually mean that the effects of the treatment *policy* are tested, rather than the effects of the treatment *per se* (Lachin 2000). But, when a threat such as having incomplete blocks is

TABLE 21.1. Matrix of possible treatment pairings in a 12-participant X 4 treatments X 2-treatments-per-block study

Treatment type	Trial blocks					
	1	2	3	4	5	6
I	X	X	X			
II	X			X	X	
III		X		X		X
IV			X		X	X

known in advance, the overall design can and in fact should be modified. Otherwise, there is the risk that the analysis will be misleading.

One of the solutions suggested in the literature, which dates back to Yates (1936), is to use a balanced *in*complete block design, or BIBD (see Federer and Nguyen 2002; Fisher and Yates 1963; Milliken and Johnson 1996; Robinson 1972). In a BIBD, not all treatments occur in every block. However, all pairs of treatments (e.g., intervention and nonintervention) occur equally often in the same block. In this way, there are pairs of treatments that the researcher can compare: hence the "balance." The trick is to create enough blocks, so that all combination pairs will occur and thus be compared. Therefore, the number of blocks that are necessary for creating this treatment balance will be determined based on the number of treatments which can be administered in a single block.

The balance is guaranteed by the following procedure. We denote a as the number of treatments, and r as the number of times all treatments are tested in the trial. Therefore, $N = (a) \times (r)$. k is used to indicate the number of treatments that occur per block, and b is the number of blocks. Therefore, b can be calculated by N/k. Put together, these parameters should satisfy $(b) \times (k) = (a) \times (r)$.

Assume for example that there are 12 observations in a trial, with 4 possible treatments throughout the study. However, we know that only 2 treatments can occur in each block (that is, $k = 2$) because it would be too expensive, for example, to have all treatments running in all blocks. Therefore, this design is incomplete. The number of times each treatment should be run in the trial can be extracted by converting the above formula, such that $r = N/a$. Because there are 12 observations and each 2 treatments can occur only once in each block, then $b = 6$ $(=12/2)$. The number of pairs in each block of our hypothetical trial is 1, as signified by "λ," which is extracted using the following formula (see Ryser 1963):

$$\lambda = \frac{r(k-1)}{(a-1)}$$

Thus, the only possible matrix, in which each pair of treatments appears simultaneously within a block exactly once, is the one presented in Table 21.1. Each pairs of treatments is then analyzed.

PERMUTED BLOCK RANDOMIZATIONS. The two designs presented earlier (complete and incomplete) aim to reduce variance in the data caused by certain prognostic or other general variables. However, these models are less useful to address time-related biases created by sequential assignment of units. There are times when the researcher needs to achieve near-equality in the number of units assigned in each block to different treatments, at any stage of the recruitment process, not only at the end of the study. A special type of block randomized design is called for.

Suppose that a care-provider cannot wait for more victims of a similar type to surface, in order to provide victims with much-needed services at the same time. The treatment must be delivered as soon as possible for each individual victim. Such is also the case in deferred enrolment designs, where participants may be rejected at one point but then enrolled at a different point in time when more appropriate participants surface for a particular treatment (e.g., finish their prison sentence, become legally fit for a rehabilitation program, graduate from a training program, etc.). However, when using a CRD to evaluate this or similar treatments, the random assignment procedure could be threatened with several possible biases, such as selection bias and chronological bias. These and other threats, which stem from changes that occur in participants' characteristics across time (or even from chance variation alone), are more likely to take place in a complete randomization design (Devereaux et al. 2005; Kao et al. 2007: 364).

First introduced by Hill (1951; see Armitage 2003), the permuted-block randomization design, or PRBD, was developed in order to deal with these trial conditions. Over the years, PRBD has become the most frequently employed design in clinical trials (Abou-El-Fotouh 1976; Cochran and Cox 1957; Lagakos and Pocock 1984; Matts and Lachin 1988; Rosenberger and Lachin 2002: 154). However, in criminology, it is quite rare (Ariel 2009).

Consider the following permutational sequence with four participants in each block (Beller et al. 2002). Here, this means that the block size, or length, is 4, with one treatment group (T) and one control group (C). Deductively, this means that in each block, there are only six possible arrangements of the treatment and control: CCTT, TTCC, CTTC, TCCT, TCTC, and CTCT. Randomization is applied in order to select a particular block out of the six possible combinations, therefore setting the allocation sequence for the first four participants. This process is then repeated, depending on the number of participants participating in the study.

The PRBD is most appropriate when there is a strong need for both periodic and final balance in the number of participants assigned to each study group in each block. This is the case, for example, in small studies, or studies with many small subgroups (Rosenberger and Lachin 2002). It is therefore useful when the study incorporates interim analyses, as in studies with increased time lags between one recruitment stage and the next, or with a longitudinal design (Kernan et al. 1999). This deals with the problem which was best described as the "time-heterogeneity bias within blocks," created by changes that occur in participants' characteristics and responses with time of entry into the trial (see Rosenberger and Lachin 2002: 41–45). The PRBD model works to correct for this (Matts and Lachin 1988).

A Word of Caution. One of the drawbacks of using a PRBD is the risk caused by the very nature of the design, namely "forcing equality" within the blocks (Schulz and Grimes 2002; see also Efron 1971; Lachin 1988b). Creating a known combination of allocation in advance jeopardizes the unpredictability of treatment assignments, which in turn may lead to intentional or unintentional unmasking of the allocation sequence, as well as prediction of future allocation, even though the blocks are selected at random.

When the size of the block is fixed and known, and the study is not double-blind (which is usually the case), this allows unmasking. It is caused by the imposition of periodic balance at the end of each successive block (Berger 2005a; Doig and Simpson 2005; Proschan 1994). Since the allocation process as well as administration of the treatment become constant over time, it is possible to know to which group the next participant will be allocated. This is particularly the case when the block size is small, with "six participants or less" (Schulz and Grimes 2002).

One worrying outcome of knowing future allocation is the increased likelihood of a selection bias. This bias implies that the outcome differences may be explained as a result of something else other than the treatment per se. While selection bias exists in all types of trials, it is more of an issue in permuted block randomization designs because, as stated, the random assignment structure is predetermined. It can lead to artificially low p values, artificially large estimated magnitudes of benefit, and artificially narrow confidence intervals (Berger 2006). The overall validity of the conclusions is also jeopardized, as well as the integrity of the trial (and possibly that of the experimenter). Therefore, the best way to decrease the likelihood of this bias would be to keep the researcher(s) implementing the trial and those controlling the random allocation sequence separate.

A CLOSER LOOK AT THE BLOCKING CRITERION

As we have tried to show, blocking on key variables is expected to create subgroups of participants which are more similar to one another than in the generally heterogeneous random sample. Whichever type of RBD is being used, this criterion can be established based on either the characteristics of the experiment (e.g., time, measurement instrument, consignment, etc.), or those of the participants (e.g., income, age, criminal background, psychological variables, or even the level of the dependent variable at pre-test, baseline level etc.). The variance between participants within blocks is then expected to become smaller than the overall variance (Abou-El-Fotouh 1976; Armitage 2003: 926; Fisher 1935). When the data are properly blocked, the estimates of the treatment effect will become more accurate because the overall power of the test statistic is increased, thus enabling the researcher to detect smaller differences between treatment means (Ostle and Malone 2000; Rosenberger and Lachin 2002).

In certain cases, it is immediately obvious that blocking should be utilized for the purposes of reducing experimental error resulting from lower variance. Blocking the data according to type of drug use, as in the hypothetical drug treatment trial presented above, is very clear. But there are other instances when the advantages of blocking are not so obvious.

Because the decision to block the data is usually based on qualitative grounds,[1] there is always the risk that the researcher has made a poor decision. In a scale-level variable, for example, this means that the cutting-points of the data might be misplaced, therefore implementing a grouping criterion that creates blocks in which the units within each block do not share common characteristics. In a study of tax evasion, for example, blocking the data according to income levels in a way that is intended to divide the sample into disparate socioeconomic backgrounds could go wrong if the researcher mistakenly categorizes participants so the blocks do not contain taxpayers with similar socioeconomic attributes. Thus, if the blocking is wrong, the block design will not only be disadvantageous compared with using an ordinary complete randomization design, it could be counterproductive and increase the error rate.

[1] Recently, Haviland and Nagin (2005: 2007) have proposed group-based trajectory modelling, which tests for treatment effects within trajectory groups. Their procedure can empirically and effectively cluster together groups that have similar baseline characteristics, for the purpose of then comparing them at the postrandomization stage.

Was the Blocking Efficient?

In order to quantify the improvement of using a blocked design over a complete randomization design, Relative Efficiency (RE) is calculated (Yates 1936). RE is a way to answer the question: how much have we gained by using a blocked design rather than a complete randomization design, with the same number of experimental units? (Hinkelmann and Kempthrone 1994: 260). Formally, RE compares the variance estimate of a blocked design with a counterfactual variance estimate of a complete randomization model. The larger the RE, the more effective the blocking has been. There are different ways to measure RE. However, the best way is to look at the estimated ratio of improvement in the context of the estimation of treatment comparisons, which heavily relies on the variance of each design:

$$\text{RE } (\text{D}_1 \text{ to } \text{D}_2) = \frac{\text{Efficiency D}_1}{\text{Efficiency D}_2} = \frac{\text{Var}_{\text{D}_1}}{\text{Var}_{\text{D}_2}}$$

Where D_1 and D_2 are the two designs.

Of course, the *true* ratio is not known, because we only have the true variance for the blocked design that is found in the observed data, not that which exists in the counterfactual CRD. Therefore, we use an *estimated* RE. The sources of data for the RE formula are taken from the mean square of the error term and the blocking term. Using the same terms discussed thus far (a being the treatment factor; b representing the blocking factor), Hinkelmann and Kempthrone (1994) suggest the following equation, referred to as the Expected Mean Squares for Treatment and Error approach, as a way to calculate RE:

$$\text{RE } (\text{D}_1 \text{ to } \text{D}_2) = \frac{(b-1)(\text{MS}_{\text{block}}) + b(a-1)\text{MS}_{\text{error}}}{(ba-1)\text{MS}_{\text{error}}}$$

In the numerator, the degrees of freedom of the blocking factor $(b-1)$ are multiplied by their respective mean squares, which are then added to the number of blocks used, multiplied by the product of the degrees of freedom of the treatment factor and the mean squares of the error term. In the denominator, we see that the mean squares of the error term are multiplied by the degrees of freedom of the total variance. If the ratio is larger than 1, we can say that the blocking factor was efficient. Hinkelmann and Kempthrone (1994) further suggest a criterion, where it may be advisable to consider a blocked design with RE larger than 1.25 as "better" than the comparable, counterfactual complete randomization model (p. 262). The question remains, however, what to do when RE is smaller than 1, or when it is clear from the data that blocking was counterproductive?

The Case of Unsuccessful Blocking

One way to understand the problem of poor blocking is to observe the relationship between the error term of the model and the degrees of freedom (see Ahamad 1967; Weisburd and Taxman 2000). As we explained earlier, the residual variance includes the degrees of freedom which were unaccounted by the model. These degrees of freedom stem from both random as well as systematic residual variance. In a nonblocked design, the residual variance term will include both the ordinary unexplained variance as well as the blocking variance. Likewise, the degrees of freedom that "belong" to the systematic variance that actually exists between

blocks of data are also located in this residual variance error term. Therefore, as Cochran and Cox (1957) described it, "blocking takes away degrees of freedom from the estimate of the error variation." This is why, as briefly discussed above, poor blocking translates into lower precision in the estimate of error variability.

An operational dilemma soon emerges: when in reality, there are no real differences between the blocks – or when RE is smaller than 1 – how should the researcher best deal with the degrees of freedom of the blocking factor in the statistical analysis? One approach argues that these degrees of freedom should *not* be included in the residual variance because it rewards the researcher who has ineffectively blocked the sample. Therefore, the blocking variable, with its respective degrees of freedom, must be taken into account. Under this approach, conservative statisticians would most likely suggest that, once an experiment has been conducted in a RBD, the blocks cannot be ignored in the analysis and the degrees of freedom should not be included in the residual variance term (see in Devereaux et al. 2005; Hinkelmann and Kempthrone 1994: 260; see review in Lachin 1988a).

However, we believe that this conservative approach is too stringent. A more practical approach to deal with this situation would be to ignore the blocking factor and analyze the data, *ex ante,* without it. We are not alone in this approach. Matt and Lachin (1988b) suggest that, when the blocking factor is ignored, the degrees of freedom would then go into the residual variance term. Disregarding the blocking effect is acceptable, because ignoring this factor should only result in a more conservative statistical test (Friedman et al. 1985). In turn, pooling the degrees of freedom may reward the researcher who has ineffectively blocked the data, but the overall variance in a poorly blocked study should in theory be equal to the overall variance of an unblocked study (Matts and Lachin 1988).

Thus, ignoring the blocking factor in the analysis is likely to result in similar outcomes as if the blocking was considered – particularly when the participants were sampled from a homogeneous population (Lachin et al. 1988). Analyzing the data this way will "simply" mean that we sacrifice both power and precision, but not the overall integrity of the study as long as the researcher implicitly reports this procedure. (Note, that this dilemma and how to solve it is only raised in relation to a design that has no intrablock treatment replication. In more complex designs, which allow for such replication, then the analysis *must* take into account the interaction between the treatment factor and the blocking factor. Therefore, the blocking criterion cannot be ignored in these cases).

Blocking *Ex Ante* Vs. Blocking *Ex Post*

The last issue we wish to consider in regard to these designs is whether prerandom blocking is at all necessary. As we have tried to show, one of the major objectives of RBDs is to decrease the variance of the data, by subdividing the sample according to key criteria. In this way, the researcher can decrease the Type I error rate as well as increase precision and statistical power. This is achieved by the prerandomization blocking procedure. However, ordinary post hoc subgroup analyses, that are generally used in analyses of variance procedures (such as Tukey's *HSD*, *Scheffe*, *Bonferroni* and the like), are used for the same reason. These analyses allow the researcher to evaluate the treatment effect on particular groups or subgroups of participants. But in this context, they can also be used to theoretically homogenize the data according to certain key variables, much like a blocking procedure. We have covered quite a few operational as well as statistical problems in RBDs, particularly when dealing with an incomplete block. Therefore, should the researcher implement an ordinary complete

randomization design instead of the relatively complex blocked design, and deal with the categorization of the data at the post hoc stage of the statistical analyses? Should certain covariates be treated *ex post* instead of *ex ante*?

Subgroup analyses are customarily viewed as a natural step that comes after testing for main effects. These analyses can provide valuable information about both planned and unanticipated benefits and hazards of the intervention. At the very least, researchers use them in order to establish treatment benefits in subsets of participants. It makes sense to first assess the treatment by comparing *all* experimental units with *all* control units (or before–after, depending on the design) and then to account statistically for any covariates or other baseline variables. It seems that most clinicians agree with this rationale as 70% of clinical trial reports include treatment outcome comparisons for participants subdivided by baseline characteristics at the postrandomization stage (Assmann et al. 2000).

At the same time, however, as logical as subgroup analyses may be in theory, it is well established now that they are often misleading (Lee et al. 1980; Moye and Deswal 2001; Peto et al. 1995). In fact, the medical community often rejects such findings (Moye and Deswal 2001), as the methods and procedures implemented are commonly misused (Assmann et al. 2000). Among some of the concerns raised against subgroup analyses, Moye and Deswal (2001) emphasize that "lack of prospective specification, inadequate sample size, inability to maintain power, and the cumulative effect of sampling error" complicate their interpretation. Some go as far as saying that the most reliable estimate of the treatment effect for a particular subgroup is the *overall* effect rather than the observed effect in that particular group (Schulz and Grimes 2005). Therefore, using subgroup analysis instead of blocking of the data before random assignment may actually be ill-advised.

Moreover, because virtually any covariate can be used to cluster units into subgroups, it allows, at least from a technical standpoint, the ability to generate multiple comparisons. This is a source of concern, because it increases the probability of detecting differences simply by chance alone (Type I error). It is not uncommon for researchers to be tempted to look for statistically significant, often publishable, differences between subgroups. It is therefore recommended that, without proper planning and sound rationale for conducting the analysis, subgroup analyses should not be considered as a replacement for prerandomization blocking. Subgroup analyses should be justified on theoretical grounds *a priori*, in order to avoid the appearance of improper data-mining. As described by Weisburd and Britt (2007: 320), "this is a bit like going fishing for a statistically significant result. However, sometimes one or another of the pairwise comparisons is of particular interest. Such an interest should be determined before you develop your analysis… if you do start off with a strong hypothesis for a pairwise comparison, it is acceptable to examine it, irrespective of the outcomes of the larger test. In such circumstances, it is also acceptable to use a simple two-sample t-test to examine group differences." Schulz and Grimes (2005: 1658) further develop this argument, by emphasizing that "seeking positive subgroup effects (data-dredging), in the absence of overall effects, could fuel much of this activity. If enough subgroups are tested, false-positive results will arise by chance alone… Similarly, in a trial with a clear overall effect, subgroup testing can produce false-negative results due to chance and lack of power."

In summary, subgroup analyses can be misleading because they focus on the heterogeneity of the intervention effect among the blocks. However, this heterogeneity is not always the question that the study is addressing (Yusuf et al. 1991). Ordinarily, hypotheses tested in the trial investigate an overall direction of the treatment effect in the study population, whereas there is no assumption of homogeneity of effect across subgroups. As shown by Adams (1998: 770), subgroups are created from the original study population, *ex post*, which

may have unknown imbalances in baseline characteristics (e.g., risk factors) that cannot be adjusted for and can influence outcomes outcome. Thus, subgroup analyses oftentimes do not reveal the truth about the relationship in the larger population. The interpretation of their results should be seen as exploratory because they can suggest but not confirm a relationship in the population at large (Moye and Deswal 2001).

Despite the reservations that we just reviewed, subgroup analyses should not be completely neglected. There are times when they can in fact replace prerandomization blocking designs. This, however, should be done with caution. When the researcher specifies at the beginning of the trial that the efficacy of the treatment for a particular subgroup or a particular block is of particular interest and part of the research goals, then subgroup analyses can be considered. When this is the case, there are good strategies that can be used to estimate the effect size in subgroups created after random assignment (see Moye and Deswal 2001). These are strategies that rely primarily on the use of prospective devices to improve safeguard from sampling error. At the same time – and however promising – they are quite underused in medical research (Assmann et al. 2000) and are rarely used in criminal justice trials. The threats which arise in such an approach may somewhat cancel out the benefits.

CONCLUSIONS

Under certain conditions, RBDs are more useful when compared with ordinary completely randomized allocation procedures. They allow the researcher flexibility and control over the number of conditions assigned to the participants, as well as the number of blocks that are used to homogenize the data. By reducing the intrablock variance, the treatment estimates are more accurate because of the increased statistical power and precision of the test statistics.

The various statistical models designed to accommodate the different types of blocking techniques provide the researcher with accuracy that is generally superior to that which can be obtained using the non-blocked randomization designs. This is particularly the case when the trial is small – several hundred or less – or when the blocking criterion is "good." Unlike the complete randomization design, these blocking procedures can deal with certain shortcomings that cannot be eliminated by randomization theory: increased data variance, outliers, time heterogeneity due to sequential assignment, missing data and covariance imbalance.

In this chapter, we reviewed some common types of the randomized blocked designs. Perhaps deterred by their complexity, the majority of criminologists have failed to use these designs. However, there are instances when they are better equipped to explore the treatment effect. *Post hoc* subgroup analyses, which are generally utilized in the simple randomization model, are for the most part not good replacements for prerandomization allocation through blocking. Blocking procedures create a more powerful treatment estimate – not by altering the effect of the treatment, which is held constant – but by decreasing the variance. This leads to more efficient estimates. Thus, these designs should be implemented in criminology more fully in the future.

Figure 21.3 aims to assist in identifying the best randomization design (CRD or RBD) for RCTs in criminology. Before selecting a complete randomization design or a RBD, the researcher should answer four interconnected questions (that appear in Fig. 21.3 as diamonds at the top of the page), all of which should be addressed at the planning stage of the experiment:

1. Is the number of participants expected to be assigned to treatment groups smaller than "a few hundred"?

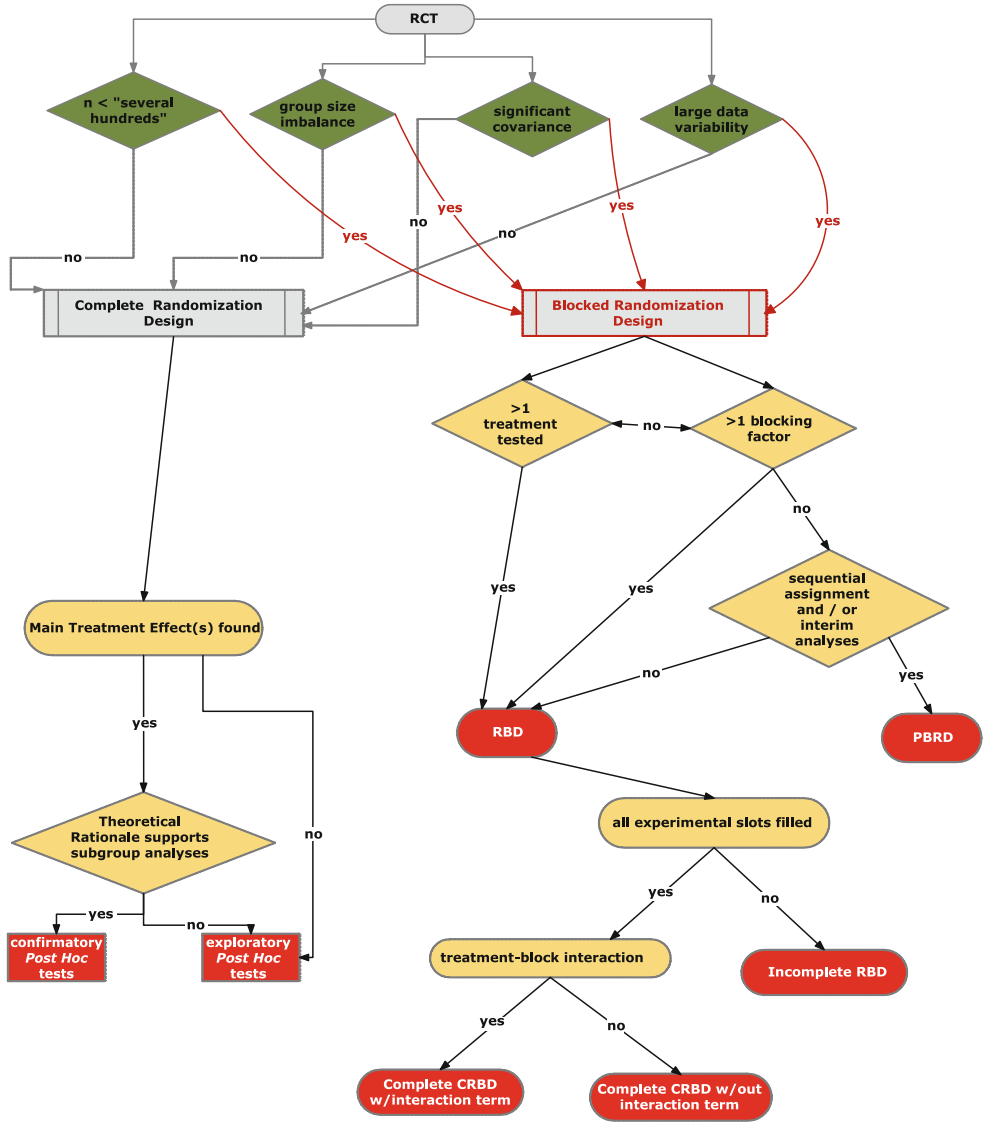

FIGURE 21.3. Identifying the best randomization design to fit the trial.

2. Is an imbalance in group-size expected?
3. Are there known covariates that are likely to have an effect on the data?
4. Are the data characterized by variance, especially that which will have a systematic impact on the distribution of the responses to the treatment?

If the answer to this entire set of questions is no, then the researcher is likely to benefit from a simple randomization design. In such a case, the treatment(s) effect should be analyzed using the statistical model which is the most useful for the nature of the data collected in the course of the experiment. *Post hoc* subgroup analyses can then be conducted, depending on whether there is a good *a priori* theoretical reasoning to justify this analysis. If not, or when

main effects were not detected, then the findings of such *post hoc* tests should be viewed as exploratory, helping to generate hypotheses for future empirical exploration. Alternatively, when there is a good reason to do them, then *post hoc* tests may be considered confirmatory, but only under certain conditions.

However, even if only one of the four questions is answered yes, this would provide justification for using a RBD. In this case, if there is only one blocking factor and only one treatment, and the researcher anticipates sequential assignment, then the design which is likely to be most useful will be the permuted block randomization design. Otherwise, the more general simple RBD is better.

The next step would be to ask whether all experimental slots are likely to be filled. This is particularly relevant when there are several treatments tested together, or when it will be too expensive to have every single slot in every block filled. In cases where all slots are not filled, the complete RBD should be used. In the more likely scenario, where every slot will in fact be filled, the researcher should then ask whether an interaction is expected between the blocking variable and the treatment variable. If so, a CRBD with an interaction term should be used. If not, a CRBD without an interaction term model should be preferred. Either way, the statistical analysis, as presented above, is usually straightforward and supported by most statistical software packages. For the most part analysis of variance procedures are most common. Other test statistics can be better, depending on the type of outcome measures assessed though these are beyond the scope of this chapter.

There have been a few RBDs in criminology, and researchers have learned about how to use the simplest type of RBDs. The time is ripe for researchers to experiment with more complicated designs, building on knowledge gained in medical research.

REFERENCES

Abou-El-Fotouh HA (1976) Relative efficiency of the randomized complete block design. Exp Agric 12:145–149

Adams K (1998) Post hoc subgroup analysis and the truth of a clinical trial. Am Heart J 136(5):753–758

Ahamad B (1967) An analysis of crimes by the method of principal components. Appl Stat 16(1):17–35

Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, Gøtzsche PC, Lang T (2001) The revised CONSORT statement for reporting randomized trials: explanation and elaboration. Ann Intern Med 134: 663–694

Ariel B (2008) The effect of written notices on taxpayers' reporting behavior – a blocked randomized controlled trial. Presented at the third annual conference on randomized controlled trials in the social sciences: methods and synthesis, York University, UK (October)

Ariel B (2009) Systematic review of baseline imbalances in randomized controlled trials in criminology. Presented at the communicating complex statistical evidence conference, University of Cambridge, UK (January)

Armitage P (2003) Fisher, Bradford Hill, and randomization: a symposium. Int J Epidemiol 32:925–928

Assmann S, Pocock S, Enos L, Kasten L (2000) Subgroup analysis and other (mis)uses of baseline data in clinical trials. Lancet 355(9209):1064–1069

Beller EM, Gebski V, Keech AC (2002) Randomization in clinical trials. Med J Aust 177:565–567

Berger VW, Exner DV (1999) Detecting selection bias in randomized clinical trials. Control Clin Trials 20:319–327

Berger VW (2005a) Is allocation concealment a binary phenomenon? Med J Aust 183(3):165

Berger VW (2006) Varying the block size does not conceal the allocation. J Crit Care 21(2):299

Braga A, Weisburd D, Waring E, Mazerolle LG, Spelman W, Gajewski F (1999) Problem-oriented policing in violent crime places: a randomized controlled experiment. Criminology 37:541–580

Canavos G, Koutrouvelis J (2008) Introduction to the design and analysis of experiments. Prentice Hall, Elk Grove Village, IL

Chow S-C, Liu J-P (2004) Design and analysis of clinical trials: concepts and methodologies. Wiley-IEEE, Taiwan

Cochran WG, Cox GM (1957) Experimental designs. Wiley, New York

Dean A, Voss D (1999). Design and analysis of experiments. Springer Science, New York

Devereaux PJ, Bhandari M, Clarke M, Montori VM, Cook DJ, Yusuf S, Sackett DL, Cina CS, Walter SD, Haynes B, Schunemann HJ, Norman GR, Guyatt GH (2005) Need for expertise based randomized controlled trials. Br Med J 7482:330–388

Doig GS, Simpson F (2005) Randomization and allocation concealment: a practical guide for researchers. J Crit Care 20:187–191

Efron B (1971) Forcing a sequential experiment to be balanced. Biometrika 58:403–417

Farrington DP, Ttofi MM (2009) Reducing school bullying: Evidence-based implications for policy. In M. Tonry (Ed.) Crime and Justice 38:281–345. Chicago: University of Chicago press

Federer W, Nguyen N-K (2002) Incomplete block designs. In: El-Shaarawi A, Piegorsch W (eds) Encyclopedia of environmetrics, Vol. 2. Wiley, Chichester, pp 1039–1042

Fisher RA (1935). The design of experiments. Oliver and Boyd, Edinburgh

Fisher RA, Yates F (1963) Statistical table for biological agricultural and medical research, 6th edn. Hafner, New York

Friedman LM, Furberg CD, DeMets DL (1985) Fundamentals in clinical trials, 2nd edn. PSG Publishing Company, Littleton, MA

Gacula M (2005) Design & analysis of sensory optimization. Blackwell, Australia

Gill JL (1984) Heterogeneity of variance in randomized block experiments. J Anim Sci 59(5):1339–1344

Hallstrom A, Davis K (1988) Imbalance in treatment assignments in stratified blocked randomization. Control Clin Trials 9(4):375–382

Haviland MA, Nagin SD (2005) Causal inference with group-based trajectory models. Psychometrika 70:1–22

Haviland MA, Nagin SD (2007) Using group-based trajectory modeling in conjunction with propensity scores to improve balance. J Exp Criminol 3:65–82

Hill AB (1951) The clinical trial. Br Med Bull 7:278–282

Hinkelmann K, Kempthrone O (1994) Design and analysis of experiments: introduction to experimental design. Wiley, New York

Hoshmand R (2006) Design of experiments for agriculture and the natural sciences. Chapman & Hall, Florida

Kao L, Tyson J, Blakely M, Lally K (2007) Clinical research methodology I: introduction to randomized trials. J Am Coll Surg 206(2):361–369

Kepner J, Wackerly D (1996) On rank transformation techniques for balanced incomplete repeated-measures designs. J Am Stat Assoc 91(436):1619–1625

Kernan WN, Viscoli CM, Makuch RW, Brass LM, Horwitz RI (1999) Stratified randomization for clinical trials. J Clin Epidemiol 52:19–26

Lachin JM (1988a) Properties of simple randomization in clinical trials. Control Clin Trials 9:312–326

Lachin JM (1988b) Statistical properties of randomization in clinical trials. Control Clin Trials 9:289–311

Lachin JM (2000) Statistical considerations in the intent-to-treat principle. Control Clin Trials 21(3):167–189

Lachin JM, Matts JP, Wei LJ (1988) Randomization in clinical trials: conclusions and recommendations. Control Clin Trials 9:365–374

Lagakos SW, Pocock SJ (1984) Randomization and stratification in cancer clinical trials: An international survey. In: Buyse ME, Staquet MJ, Sylvester RJ (eds) Cancer clinical trials, methods and practice. Oxford University Press, New York, pp 276–286

Lee KL, McNeer F, Starmer CF, Harris PJ, Rosari RA (1980) Clinical judgment and statistics: lessons from a simulated randomized trial in coronary artery disease. Circulation 61:508–515

Liebetrau A (1983) Measures of association. Sage, Thousand Oaks, CA

Lipsey MW (1990) Design sensitivity: statistical power for experimental research, Sage, Newbury Park, CA

Matts JP, Lachin JM (1988) Properties of permuted-block randomization in clinical trials. Control Clin Trials 9:327–344

Milliken AG, Johnson DE (1996) Analysis of messy data. Van Nostrand Reinhold, New York

Mottonen J, Husler J, Oja H (2003) Multivariate nonparametric tests in a randomized complete block design. J Multivar Anal 85:106–129

Moye L, Deswal A (2001) Trials within trials: confirmatory subgroup analyses in controlled clinical experiments. Control Clin Trials 22(6):605–619

Ostle B, Malone L (2000) Statistics in research: basic concepts and techniques for research workers. Iowa State University Press, Wiley-Blackwell

Palta M, Amini SB (1982) Magnitude and likelihood of loss resulting from non-stratified randomization. Stat Med 1(3):267–275

Peto R, Collins R, Gray R (1995) Large-scale randomized evidence: large, simple trials and overviews of trials. J Clin Epidemiol 48:23–40

Pocock SJ, Simon R (1975) Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. Biometrics 31:103–115

Pocock SJ (1979) Allocation of patients to treatment in clinical trials. Biometrics 35:183–197

Proschan M (1994) Influence of selection bias on type i error rate under random permuted block designs. Stat Sin 4:219–231

Robinson J (1972) The randomization model for incomplete block designs. Ann Math Stat 43(2):480–489

Rosenberger W, Lachin JM (2002) Randomization in clinical trials: theory and practice, Wiley, New York

Ryser HJ (1963) Combinatorial Mathematics. Cambridge Mathematical Monographs No. 14. The Mathematical Association of America. John Wiley and Sons.

Schulz KF, Grimes DA (2002) Generation of allocation sequences in randomized trials: chance, not choice. Lancet 359:515–519

Schulz KF, Grimes D (2005) Multiplicity in randomized trials II: subgroup and interim analyses. Lancet 365(9471):1657–1661

Senn JS (1989) Covariate imbalance and random allocation in clinical trial. Stat Med 8:467–475

Senn JS (1994) Testing for baseline balance in clinical trials. Stat Med 13(17):1715–1726

Sherman WL (1993) Defiance, deterrence, and irrelevance: a theory of the criminal sanction. J Res Crime Delinq 30:445–473

Sherman L, Weisburd D (1995) General deterrent effects of police patrol in crime hot spots: a randomized controlled trial. Justice Q 12:625–648

Sherman L, Gartin P, Buerger M (1989) Hot spots of predatory crime: routine activities and the criminology of place. Criminology 27:27–56

Simon R (1979) Restricted randomization designs in clinical trials. Biometrics 35:503–512

Torgerson JD, Torgerson CJ (2003) Avoiding bias in randomized controlled trials in educational research. Br J Educ Stud 51(1):36–45

Wei L, Zhang J (2001) Analysis of data with imbalance in the baseline outcome variable for randomized clinical trials. Drug Inf J 35:1201–1214

Weisburd D, Britt C (2007) Statistics in criminal justice. Springer, New York

Weisburd D, Green L (1995) Policing drug hot spots: the Jersey City DMA experiment. Justice Q 12:711–736

Weisburd D, Morris N, Ready J (2008) Risk-focused policing at places: an experimental evaluation. Justice Q 25(1):163–200

Weisburd D, Petrosino A, Mason G (1993) Design sensitivity in criminal justice experiments. Crime Justice 17(3):337–379

Weisburd D, Taxman FS (2000) Developing a multicenter randomized trial in criminology: The case of HIDTA. J Quant Criminol 16(3):315–340

Yates F (1936) A new method of arranging variety trials involving a large number of varieties. J Agric Sci 26:424–455

Yusuf S, Collins R, Peto R (1984) Why do we need some large, simple randomized trials? Stat Med 3(4):409–420

Yusuf S, Wittes J, Probstfield J, Taylor HA (1991) Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. J Am Med Assoc 266:93–98