4

Methodological Fundamentals for RETs

You can't fix by analysis what you bungled by design

- LIGHT, SINGER AND WILLIT

INTRODUCTION

A great deal has been written about optimal methodological practices for randomized trials (Elbridge & Kerry, 2012; Friedman, Furberg, DeMets, Reboussin & Granger, 2015; Solomon, Cavanaugh & Draine, 2009). In this chapter, I provide a brief review of these practices as they apply to RETs. My treatment is selective; to do otherwise would require an entire book on RCT design. I begin by describing different types of randomized trials, including parallel-group trials, comparative trials, non-inferiority trials, and efficacy versus effectiveness trials. I then consider trial designs, including the two-group pretest-posttest design, clustered designs, wait-list designs, crossover designs, and adaptive designs. Next, I consider the concept of a population as contextualized in randomized trials and use this to develop the implications of sample imbalance in random assignment. I then describe the mechanics of randomization. I discuss the phases of randomized trials, demand characteristics and treatment integrity. Finally, I present and comment on the Consolidated Standards of Reporting Trials (CONSORT) checklist for reporting randomized trials. The material in this chapter is fundamental to good design of RETs. Not only must you have your conceptual (Chapter 2) and measurement (Chapter 3) houses in order when conducting evaluations, so too must you have strong design/methodology. In this chapter, I identify key design issues you need to consider.

TYPES OF RANDOMIZED TRIALS

There are many ways of characterizing randomized trials. One of the most common types is a **parallel groups trial**, which seeks to determine if a program or an intervention has an effect on an outcome relative to a “neutral” (parallel) condition. The neutral condition can take many forms. For example, a trial might evaluate the effects of having cigarette smokers write counter-attitudinal essays about reasons not to smoke on subsequent cigarette smoking behavior. Individuals in the treatment condition come to an office, sit in a quiet room, and are provided a cover story for why they are to write the essay. One month later, participants are contacted by phone to complete a general survey on health, with one of the questions assessing the frequency of cigarette smoking during the past month. Individuals in the control condition are contacted at the time of the phone assessment and administered the same survey. Indices of smoking behavior are then compared between the two conditions. The latter group is called a **no contact control group** or, alternatively, a **passive control group**. Alternatively, individuals in the control condition night come to the office and write a counter-attitudinal essay, but they would do so on an unrelated topic. They then complete the health survey by phone one month later. This represents an **active control group**. With an active control group, the tasks are equated as much as possible between the treatment and control groups except for the presumed “active ingredients” of the treatment program. In traditional drug trials, it is represented by a placebo condition in which individuals in the control condition are given an inactive drug (a sugar pill). The use of an active control group in behavioral research can sometimes be more costly than a passive control group.

A second type of trial is known as a **comparative trial**. These trials compare two or more treatments/interventions to one another. These designs often (but do not always) have three conditions to which people are randomized, (1) program A, (2) program B, and (3) a control condition that is either active or passive in nature. Sometimes, the control condition is omitted. This might be the case when the efficacy or effectiveness of the two programs has already been established and it is felt there is no need to include the control condition for purposes of documenting program effects. Rather, the primary interest is to compare the two programs to determine their relative effectiveness. Although there will be cases where this rationale is reasonable, for RETs, it usually is better to include a control group. One reason is that the RET goal is not only to examine program effects on outcomes but also program effects on mediators. This is best accomplished through treatment-control comparisons, as I elaborate below. A second reason is that by replicating effect sizes on the outcome from prior research for each program, it reassures critics that you have faithfully implemented the programs.

Sometimes one program in a comparative trial is a new intervention and the other program is “treatment as usual” (TAU), which often is conceptualized as a control group. A TAU condition is an appropriate comparator to a new intervention when the primary interest is improving on the status quo. However, some scientists criticize the use of TAUs as control conditions when the goal is to test or advance theory. For example, two researchers might evaluate the same intervention under generally comparable circumstances but the particular TAU for one researcher might be reasonably effective given standard practice in his or her community whereas the TAU for the other researcher might be ineffective given standard practice in the community. The two studies might find differential program effects primarily because of quality differences in the TAU despite the fact the new programs themselves performed comparably in and of themselves. Without a careful analysis of what the TAU represents, it can be difficult to know what the new program is being compared to, thereby limiting knowledge gain. To be sure, we *do* learn that the new program is better than the status quo and that conclusion in and of itself might be useful for scenarios where the type of TAU studied is common. However, to build scientific theory, a reliance on the operative TAU in arbitrarily selected clinics can be limiting. Some methodologists argue that researchers conducting theory tests should instead consider creating an informative control condition in which the treatment and controls are equivalent in all respects except for the “active ingredients” in the program one seeks to evaluate. Sometimes that requires modifying TAU to make it equivalent to the new program minus the presumed active ingredients.

A third type of randomized trial also is a type of comparative trial but is has a special label, namely a **non-inferiority trial**. In a non-inferiority trial, the researcher is most interested in determining if two treatments are equivalent in terms of their effects on an outcome rather than if one treatment is superior to the other treatment. These designs became prominent when manufacturers of generic drugs sought drug approval from the Food and Drug Administration (FDA) by arguing that their generic drugs were less costly but equivalent in effectiveness to brand name counterparts, i.e., they wanted to show that the generic drugs were not inferior to brand name drugs. Ironically, such equivalence could be “demonstrated” in a traditional trial by using weak and underpowered experimental methods that were almost certain to produce statistically non-significant differences between the brand name and generic drug. The FDA introduced specialized designs to test for non-inferiority that circumvented this problem.

Yet another way randomized trials have been distinguished is in terms of their focus on efficacy versus effectiveness. An **efficacy trial** seeks to determine if an intervention affects an outcome given the intervention is properly implemented and individuals receive the “full dose” of the program, per protocol. For example, if a program has 6 sessions, an efficacy trial would ensure that participants in the treatment condition participate in all 6 sessions. An **effectiveness trial** seeks to determine if an intervention affects an outcome based on how the intervention is applied in real-world settings where patient populations and clinic variables cannot be rigorously controlled. In real world settings, patients might drop out of treatment, they might miss one or more sessions, they might not do “homework” if the program requires them to do so, and/or program staff might modify the intervention in subtle or even blatant ways. Efficacy trials are designed to minimize such occurrences and, if violations to protocol occur, researchers make analytic or methodological adjustments to correct for this. By contrast, an effectiveness trial allows the real world to “do its damage” so that one can determine how the program will fare once it is implemented in the broader community. Effectiveness trials are sometimes called **pragmatic trials**.

In effectiveness trials, assessments of *all* individuals who are randomized to the treatment and control conditions are made at the posttest, irrespective of whether individuals drop out of the program or do not fully engage the treatment protocol. By comparing outcome distributions as a function of the condition people were randomized to irrespective of treatment adherence and other real-world “noise,” one gains insights into treatment effectiveness. By contrast, in efficacy trials, the focus is only on individuals who complete the treatment per protocol. The former analytic approach is called **intent-to-treat** (ITT) analysis and the latter is called **per-protocol** (PP) analysis.

In a traditional randomized trial, effectiveness designs and ITT analyses are seen as "better" or "more desirable" by many methodologists. For an RET, however, the choice of an efficacy versus effectiveness orientation is complicated. For RETs that focus on program development, we want to understand the mechanisms by which a program affects an outcome in an efficacy sense. Once we have such knowledge and have improved the program to maximize its potential based on the feedback gained from the efficacy oriented RET, we then address the question of how best to move the program into real-world settings. Effectiveness trials, by contrast, confound three dynamics, (1) the extent to which a program is efficacious, (b) the extent to which the program is implemented correctly in treatment settings, and (c) the extent to which patients adhere to treatment protocols. When an effectiveness trial fails to produce meaningful outcome, we do not which of these three facets has failed us. Efficacy trials, by contrast, focus on the mechanisms that underlie outcome change devoid of the "noise" produced by improper implementation and protocol non-adherence. Program efficacy, program implementation, and protocol adherence can be influenced by different factors, so the relevant mediators and moderators for an efficacy RET can differ from those for an effectiveness RET. An effectiveness RET must address mediators and moderators for all three dynamics (efficacy, implementation, and adherence) whereas an efficacy RET need only focus on mechanisms affecting efficacy. Assessing mediators and moderators for the three different dynamics in a single RET can be challenging.

When developing new, evidence-based programs, scientists often pursue the three facets in sequenced RETs that culminate at the final stage in an effectiveness trial. The first step is to conduct an efficacy RET to better understand the mechanisms/mediators that produce change in the target outcome. For example, after an efficacy RET, one might learn that the program affects only two of the four targeted mechanisms it sought to change and that it therefore needs to be strengthened to better address the two mechanisms it failed to change. One also might learn that one of the four targeted mechanisms/mediators is, contrary to program assumptions, unrelated to the outcome. Program activities directed at that mediator might therefore be dropped. Once the program is revised based on this information, attention then turns to developing strategies for "rolling out" the revised program in real world clinic settings in ways that promote proper program implementation. These implementation strategies can be evaluated in an implementation RET where the outcome shifts to that of proper implementation. Based on this RET, the original program might be revised to make it more amenable to faithful implementation. Next, the twice-revised program is subjected to an adherence RET in which the outcome shifts to patient adherence to program protocols, working with mediators and moderators of such adherence. After the program is revised a third time to maximize client adherence, an effectiveness RET is pursued to evaluate the potential of the finished product in real world settings.

Some social scientists criticize the sequenced approach because it takes so long to execute the series of studies. Treatments can become outdated by the time the sequence is finished. In response, the National Institute of Mental Health (NIMH) has advocated for hybrid designs that combine the ability to evaluate efficacy and effectiveness simultaneously. Such designs are complex because their goals are often oppositional; on the one hand, you want to maximize implementation fidelity and adherence to accurately assess efficacy; on the other hand, you want to let fidelity and adherence vary as it would in the natural world so you can determine their effects.

When you are hired by a client to evaluate a program, the client typically is interested in the effectiveness of a program in the setting in which the program is administered. It is important to keep this focus front and center as you perform your evaluation. However, as noted, a key component of effectiveness is the efficacy of the program and you will want to gain perspectives on efficacy as well. In this sense, program evaluations for clients also require hybrid designs that can provide perspectives on both efficacy and effectiveness.

Issues surrounding the use of efficacy versus effectiveness RETs and the analytic strategies for them are complex enough that I devote an entire chapter to the matter (Chapter XX). The bottom line is that when designing an RET, you need to decide what your focus will be, namely on efficacy, implementation fidelity, protocol adherence, or some combination of the three. You then design your RET accordingly. If the decision is to focus just on efficacy, then you adopt methods to maximize fidelity and adherence, even if the ways you do so may not be realistic in real world settings. For example, you can't study the potential of a drug to cure a disease and the biological mechanisms by which that drug impacts the disease by including people in the treatment condition who take insufficient dosages of the drug or who do not take it at all. You adopt methodologies that ensure fidelity and adherence. If the decision is to focus on all three facets, then you would ensure your RET addresses mediators and moderators for each of the three facets. Your approach to RET design depends on your priorities.

RANDOMIZED TRIAL DESIGNS

The variety of randomized trial designs is considerable. I focus here on designs that I consider in future chapters (with a few exceptions), but the list of design types I provide here is not exhaustive. I first describe the classic two-group, pretest-posttest design and use it to establish the advantages of a randomized trial relative to a single group pretest-posttest design. In the context of doing so, I identify some common practices that are, in my opinion, ill-advised. I then discuss clustered designs, wait-list designs, cross-over designs, and adaptive designs.

Trial Design 1: Two-Group, Pretest-Posttest Designs

Campbell and Stanley (1963) identify factors that can undermine inferences about program effects in a variety of experimental designs. They use what they call a **single group, pretest-posttest design** to illustrate many of the validity threats we need to be concerned with, which is signified by O_1 X O_2, where the O represent the points in time when an outcome is measured and X is the administration of the intervention. Another design they discuss, and one that is quite common, is the **two-group, pretest-posttest design**. It appears diagrammatically as follows, with random assignment to the groups:

Treatment Group:	O_1 X O_2
Control Group:	O_1 O_2

In this design, a control group that is not exposed to the intervention is added to the single group, pretest-posttest design. A related design is a **two-group, posttest only** design:

Treatment Group:	X O
Control Group:	O

which is the two-group, pretest-posttest design but with no baseline assessments.

I consider here eight "threats to validity." Appreciation of these threats is essential to the design of strong RETs. Most of the threats were introduced by Campbell and Stanley using a single group, pretest-posttest design. However, the threats are relevant to two-group randomized designs when researchers pursue answers to questions by shifting their focus away from the randomized groups and concentrate instead on just the intervention group. For example, some researchers identify moderators of treatment response by calculating pretest minus posttest change scores for individuals in the intervention condition and then correlate these change scores with individual difference variables, such as gender, ethnicity, social class, or baseline severity. If a significant correlation is observed, say, between gender and the change scores, it is concluded that gender affects (or moderates) treatment response, e.g., males in the treatment condition show more change than females in the treatment condition. Note that this approach ignores the control group and thus simulates a single group, pretest-posttest focus. To the extent there are "threats to validity" for single group, pretest-posttest designs, those threats compromise such moderator analyses.

Threat 1 to Valid Inference: Testing Effects

Testing effects occur when the act of completing a measure changes people's standing on the construct being measured. Suppose in a single group pretest-posttest design, a program to lower depression is evaluated. Study participants complete an assessment battery at baseline asking about depressive symptoms and the coping strategies they use to deal with them. By completing the measures, individuals may reflect on and think more thoroughly about the symptoms and coping strategies. It may be that these reflections, not the intervention, cause reduction in depression and better coping strategies at the posttest. Testing effects are controlled in the randomized two-group pretest-posttest design if one compares the posttest outcome scores for the intervention group with the posttest outcome scores for the control group. Testing effects should affect both groups equally, so any differences between the groups on the outcome at posttest can't be attributed to testing. Alternatively, one can use the posttest-only control group design, which eliminates the baseline, thereby removing any possibility of testing effects. This works fine for outcome-only RCTs, but it sacrifices much in RETs with mediation and moderation.

Threat 2 to Valid Inference: History Effects

History effects refer to events external to the study that may be responsible for changes in the outcome rather than the intervention. For example, in the single group pretest-posttest design, the effects of a program to decrease stress among patients over a period of 3 months might be overestimated or underestimated due to events that occur in the community or the broader geographic region, such as an improving economy leading to more family income that, in turn, reduces stress. Or, an environmental program to reduce energy consumption might be implemented at the same time that the price of energy spikes upward, causing people to use less energy. History effects can be controlled by using either the randomized two-group pretest-posttest design or the two-group posttest only design because the effects of external events should occur in both the experimental and control groups. Outcome differences between the groups, as such, cannot be attributed to history effects.

Threat 3 to Valid Inference: Instrumentation Change

Instrument change occurs when the measuring device used to assess the outcome changes over time in ways that suggest program or treatment effects may be larger or smaller than is the case. An example is the case of observer drift for an outcome measure that relies on observer reports (Smith, 1986). Observer drift occurs when observer understandings of the behavioral codes they use change over time. Observer drift can produce changes in the recorded observations even if a program has no effect. For example, what is considered to be aggressive behavior by children in a playground at baseline might be seen as less aggressive at posttest as observers become more accustomed to seeing aggressive behavior occur. For self-reports, perhaps the interpretation of the meaning of items on a self-report instrument change over time as participants become more sensitized to constructs addressed in an intervention. Mean differences between baseline and posttest assessments may not be due to the program but to the different interpretations of the items on the inventory. Instrument change usually operate for both the experimental and control groups in two-group designs, so between-group comparisons of posttest outcomes control for them (but see below for exceptions). Even if equated in the two-groups, instrument change is worrisome because the construct studied might be different than one thinks.

Threat 4 to Valid Inference: Regression to the Mean

Another validity threat is **regression to the mean**. The dynamics underlying this phenomenon are not well understood by many, so I develop the concept in depth here.

Suppose I select for study individuals with extreme scores in a population distribution, such as people who score high on a depression scale. If I measure their depression again at a later point in time - even in the absence of a treatment program - I likely will find that their scores "regress to the mean" of the original distribution, i.e., the mean for the extreme group at re-assessment will change towards the value of the original mean of the full group. There are many sources of such regression to the mean, but I focus initially on one source, measurement error. I will use an unrealistic example to make it easier to illustrate the dynamics.

Consider a group of 9 individuals who, unbeknownst to me, have the same true levels of depression, which I index on a 0 to 100 metric with higher scores indicating higher levels of depression. Table 4.1 presents their observed scores on a depression measure (Y) and their true depression scores (T) at baseline (Time 1) and posttest (Time 2). I have ordered the observed scores from highest to lowest based on their Time 1 values (note: in practice, we do not know the true scores, but assume we do here for purposes of pedagogy). Consistent with classic test theory, each observed score is an additive function of a person's true score plus random noise reflecting measurement error, such as misreading items on the inventory, being distracted, and so on. The random error is uncorrelated with the true scores because it is random. Some of the random error pushes observed scores upward and some of the random error pushes observed scores downward; again, its influence is random. Note that the error scores at Time 1 are uncorrelated with the error scores at Time 2, also because the error at both times is random; correlating one set of random numbers with another set of random numbers yields a zero correlation.

Table 4. 1: Regression to the Mean

		Time 1			Time 2		
Person	`	Y1	T1	E1	Y2	T2	E2
1		93	90	+3	93	90	+3
2		93	90	+3	87	90	-3
3		93	90	+3	90	90	0
4		90	90	0	87	90	-3
5		90	90	0	93	90	+3
6		90	90	0	90	90	0
7		87	90	-3	87	90	-3
8		87	90	-3	93	90	+3
9		87	90	-3	90	90	0

Suppose I decide to focus a treatment for depression on individuals who are most depressed, so I select the three individuals with the highest observed depression scores. Unbeknownst to me, these individuals are no different in their depression levels at baseline from any of the other individuals and their elevated scores are due solely to random error. The mean observed depression score for these three individuals at baseline is 93. I expose them to the intervention to lower depression and then measure their scores at Time 2, post-intervention. The intervention is completely ineffective and the true scores of the three individuals remain the same. This is shown in Table 4.1. Because random error at one point in time is uncorrelated with random error at another point in time, the random noise that contaminates the observed scores at Time 2 will take on different values than those that contaminate the scores at Time 1. The random errors for the highest scoring individuals at Time 1 were all positive in value (each was +3), but at Time 2, the values of the random errors are now evenly distributed across the three individuals (one is +3, the other is 0, and the third is -3). The mean of the observed scores of the three individuals at Time 2 is now 90 and it looks like, at the observed score level, that the intervention had an effect because it decreased depression from a mean of 93 to a mean of 90 for these individuals. But this result is an artifact of regression to the mean. We essentially biased our selection of participants towards individuals with positive random errors at Time 1 whose errors will not necessarily be positive at Time 2.

Regression to the mean is caused not only by measurement error, but by any variable that disrupts the correlation between two sets of scores measured at different points in time. Suppose I use a measure that is perfectly reliable and valid, Y, and measure it at two time points, Y1 and Y2. There is no measurement error in Y. Suppose further that the mean and variance of Y at both times is the same and that there is a 3-month interval between measures. Suppose I select the highest 10% of the population based on their scores on Y1. Assuming the variable is temporally dynamic, the chances are low that on the second testing occasion the exact same individuals will again be the highest 10% of the sample. To be sure, many of those who were initially selected as being in the highest 10% will again be in the highest 10% at time 2; but this will not be true in every case. Variables unrelated to Y1 that are operating in the real world (i.e., disturbance variables) will have caused some people's true standing on Y2 to change upward and some people's true standing on Y2 to change downward, essentially mimicking the random influences on Y in our measurement error example. If even just a few of the originally selected people are no longer in the highest 10% because of these random disturbances, the Y2 group mean for the highest 10% will now be closer to the original population mean, which is unchanged over time. Essentially, the disturbance variables, not measurement error, cause regression to the mean. As such, regression to the mean is caused by the presence of any disturbance variable that causes the correlation between Y1 and Y2 to be less than 1.00 (see Kenny & Campbell, 1999, for elaboration). This same phenomenon operates at the lower end of the distribution in a mirror fashion, with those in the lowest 10% of scores showing improvement in their mean Y over time.

There are two key points to keep in mind. First, regression to the mean is a group phenomenon not an individual phenomenon. We usually have no idea if a given individual is going to change upward or downward over time. But, as a group, the mean of the most extreme individuals at one or the other end of the distribution is likely to move closer to the original population mean given the operation of random disturbances. Second, regression to the mean operates even if everyone exhibits true changes by a systematic amount due to some intervention or event between Time 1 and Time 2. It may be the case that everyone's true score in the population decreases by a constant of 10 units. Nevertheless, the presence of disturbance variables (such as measurement error) will cause the highest scorers at Time 1 to decrease more, on average, than the overall population decrease of 10 units and the same dynamic is mirrored for those at the lower end of the distribution but in the opposite direction.

The bottom line is that in a single group, pretest-posttest design that has selected individuals from one end of the distribution for study, regression to the mean artifacts will produce mean changes toward the full population mean. In studies evaluating interventions that rely on extreme groups (e.g., patients who are above a clinical cut-off for depression), some degree of mean reduction/change is almost assured. The best way to deal with regression to the mean is to randomly assign the (extreme) individuals to the intervention and control groups. Regression to the mean should then operate equally in both groups. Any group differences in posttest means will reflect the true intervention/treatment effect. Evaluating change within a given group (e.g., just the treatment group) is problematic because it will be subject to regression to the mean.

Threat 5 to Valid Inference: Maturation

Campbell and Stanley (1963) refer to *maturation* as “all biological or psychological processes which systematically vary with the passage of time, independent of specific external events” (pp.12-13). A single group pretest-posttest design evaluating an intervention over a six-month period with middle school adolescents, for example, is contaminated by pubertal changes in adolescents that naturally occur during this time period. If pubertal changes are relevant to one’s outcome, the evaluation study is compromised. An intervention that improves cognitive functioning in the elderly might erroneously appear to be ineffective in a single group pretest-posttest design because of naturally occurring, biologically based decrements in cognitive function that occur at the same time as the intervention. The program may increase cognitive performance but this is offset by the naturally occurring decrements that occur with aging, yielding the same pretest and posttest means. Had the program not been provided, the posttest means would have been lower than the pretest means. Maturation can be controlled using a two-group randomized pretest-posttest design because it should operate in both the intervention and control groups; any differences between them cannot be attributed to maturation.

Threat 6 to Valid Inference: Experimental Mortality

Campbell and Stanley (1963) define experimental mortality as participants dropping out of a study. In the single group pretest-posttest design, if dropping out of the study is non-random, then it is possible that biased estimates of treatment effects will occur. For example, if early responders to a treatment for depression are more apt to remain in the study but early non-responders are more apt to drop out (because the treatment does not seem to be working), then the estimated effects of the treatment will be overestimated.

Campbell and Stanley also caution about study dropouts for the randomized two-group pretest-posttest design and the randomized posttest only control group design if there is differential drop-out for the treatment and control groups. If the differential drop-out is systematic rather than random, then a treatment can appear to be more (or less) effective than it actually is. For example, a resource and time demanding treatment that is cognitively complex might lead people who are less motivated to drop out of the treatment condition than in a control condition that is less demanding. This differential drop out can bias effectiveness estimates for the treatment versus control conditions.

Threat 7 to Valid Inference: Selection Effects

Campbell and Stanley discuss selection effects in the context of two-group designs without randomization. *Selection effects* refer to self-selection into the treatment or control conditions by individuals such that the treatment and control conditions are confounded by a host of individual differences. For example, consider an after-school program to increase reading skills in students. A researcher compares post-program reading skills for youth who completed the program with students not in the program. Without random assignment, it could be that students who volunteered to be in the program have different pre-program reading skills and different motivations to achieve in school than students who did not volunteer. These pre-existing differences muddy the evaluation of program effects because one does not know if group differences in reading skills after the program are due to the program or to pre-existing group differences.

Selection effects also can undermine randomized designs if the selection variables interfere with the process of random assignment or if they exert their influence after randomization has occurred. For example, methodologists have suggested that researchers or clinic staff sometimes deviate from random assignment protocols to ensure a person who is in particular need of treatment is assigned to the intervention as opposed to the control condition (Rosenberger & Lachin, 2015). Berger (2005) suggests that researchers or clinic staff may occasionally deviate from the random assignment protocol so as to enroll patients into the intervention if they think the person is more likely to respond to treatment, thus favoring the treatment. Systematic post-randomization selection bias also can occur due to treatment dropouts, non-adherence to protocols, study attrition, missing data, and unintended between-condition differences in the use of co-occurring treatments (e.g., medications to supplement a behavioral therapy trial).

Threat 8 to Valid Inference: Interaction Effects between Threats

It is possible for any of the above phenomena to operate in interaction with each other or with treatment administration to undermine inferences about a treatment effect. For example, an intervention might only be effective if preceded by a baseline assessment because the baseline assessment sensitizes people to issues addressed in the intervention. Note that this is not the same as a testing effect per se. A testing effect refers to the impact of completing an assessment on the outcome. A testing-by-treatment interaction is when the effects of the treatment are dependent on completing the baseline assessment.

As an example of a selection-maturation interaction, suppose a study of the effect of a new teaching approach in elementary school on student math abilities is conducted, with students in the control group recruited into the study at the beginning of the school year and those in the experimental group recruited in at the end of the school year (note: this is *not* random assignment to condition). There are selection effects present such that the students in the control group are younger than those in the experimental group and one would expect increases in cognitive abilities as a result of maturation over the course of the school year, changes that could bias scores on the math ability test. Treatment versus control group differences in posttest math ability means might occur, but they likely represent a selection-maturation interaction rather than program effects.

In sum, when conducting research to test the effect of an intervention on outcomes, it is important to rule out the types of threats discussed by Campbell and Stanley (1963). Campbell and Stanley suggest an effective strategy for addressing many of these threats called a **Solomon four group design**, which uses the following randomized design:

Group 1:	O_1	X	O_2
Group 2:	O_1		O_2
Group 3:		X	O_2
Group 4:			O_2

Note that this design is a combination of the two-group pretest-posttest design (groups 1 and 2) and the single group posttest-only design (groups 3 and 4). Comparing O_2 for Group 4 with O_2 for Group 2 diagnoses testing effects. Comparing O_2 for Group 3 with O_2 for Group 1 diagnoses testing by treatment interactions. Comparing O_2 for Group 3 with O_2 for Group 4 tests for intervention/treatment effects. For more details, see Campbell and Stanley (1963) and Reichardt (2019). The Solomon four group design is rarely used because it tends to be costly.

Revisiting Within-Condition Analyses of Change Scores

I now revisit two practices mentioned earlier that you will sometimes encounter in program evaluations using randomized trials. Both practices focus on change scores within the intervention condition. The first practice is when researchers ignore the control group and test for statistically significant change in the outcome from the pretest to the posttest for the intervention group using either a mean change score (by testing if the mean change is statistically significantly non-zero), a dependent groups t test (by comparing the pretest mean to the posttest mean), or a mixed effects model, all of which are statistically equivalent and yield the same result. The tests are used to assert the program produced change in the outcome from pretest to posttest. The problem with this strategy is that the mean change does not just reflect response to treatment; it also reflects testing effects, history effects, instrumentation change, regression to the mean, maturation effects, experimental mortality, selection dynamics, and potential interactions among these confounds, as well as other factors I have not yet discussed. This is not a scientifically sound way to assert a program effect. By ignoring the control condition, it ignores the advantages of an RCT and it essentially reduces to a single group pretest-posttest, with all of its methodological problems. The best way to assert program effects is to compare the posttest mean for the treatment group with the posttest mean for the control group because the comparison accounts for these artifacts.

The second practice is when researchers seek to identify moderators of treatment response by correlating pretest minus posttest change scores with individual difference variables, again within the intervention condition only. For example, for the intervention group, I might correlate biological sex with the change scores to determine if females or males respond better to treatment. A problem with this approach is that the change scores do not only reflect response to treatment. They are confounded with testing effects, history effects, instrumentation change, regression to the mean, maturation effects, experimental mortality, and selection dynamics, among other things. A second problem with this approach is that the correlation with the change score is impacted by the degree of correlation between the individual difference variable and the outcome as measured at baseline independent of change. Let Y_{PRE} be the outcome measured at baseline and Y_{POST} be the outcome measured at posttest. A change score, CS, is Y_{PRE} - Y_{POST}. Any individual difference variable that is correlated with Y_{PRE} will exhibit a correlation with CS because Y_{PRE} is part of CS. For example, biological sex tends to be correlated with depression: Females tend to report higher levels of depression than males. Given that sex is correlated with Y_{PRE}, it also will show an artifactual correlation with the change score. Indeed, because Y_{PRE} is part of CS, it also is the case that baseline measures of the outcome will show artifactual correlations with CS (see Cohen & Cohen, 1984, for mathematical proofs). This artifact has led many researchers to erroneously conclude individuals with more extreme baseline scores respond differently to a treatment, unaware that the baseline score will be correlated with change just because it is part of the change score. A more rigorous approach to identifying moderators of treatment effects is to include both treatment and control individuals in the analysis of moderators by using a dummy variable, T, to represent the treatment versus control condition and then to formally test the interaction between T and the individual difference variable of interest (e.g., sex, ethnicity, social class) using product terms. I discuss this approach in more depth in Chapters XX and XX. Be careful not to fall into these traps.

Trial Design 2: Clustered Designs

Many trials randomly assign individuals to condition, but some trials randomly assign clusters of individuals to conditions rather than individuals per se. For example, when evaluating the effects of a new math curriculum in schools, a researcher might randomly assign classes to either the "new curriculum" or "old curriculum" condition. Or, researchers might test the utility of a new group treatment for attention deficit disorder (ADHD). They randomly assign individuals to treatment versus control conditions but they then form therapy groups of 5 individuals per group, to which a group leader administers the therapy. The treatment condition might have 50 such groups and the control condition also might have 50 groups that engage in ADHD irrelevant task activities. The therapy groups represent the clusters and this is another example of a **cluster randomized design**.

When we analyze data in a traditional randomized trial, a statistical assumption we make is that the error scores across individuals are independent. However, in clustered designs this may not be the case. For example, if a therapy group contains a particularly disruptive group member, then the outcome scores for all members of that group might be affected, but not members of other groups since they are not exposed to the disruptive individual. Or a group might have a particularly good therapist/teacher/leader and all of the members of that group benefit but not members of other groups. The presence of such cluster effects can create dependencies among the error scores. If the dependencies are strong enough, then adjustments need to be made for statistical tests to be valid.

Clusters can be small groups, they can be schools, they can be communities, or any other clustered unit that is randomly assigned to conditions or that exist within treatment and control conditions in ways that cluster membership introduces bias in statistical inference. I discuss the analysis of randomized cluster designs in Chapter 29. It is not uncommon to find examples of group-administered therapies in clinical psychology that are analyzed as if the individuals are independent when a cluster-adjusted analysis is called for.

Trial Design 3: Wait-List and Cross-Over Designs

Another type of trial design is the wait-list design. In this case, individuals are randomly assigned to either a treatment or a control condition. Individuals in the control condition receive treatment, but they are delayed in doing so in order to fulfill their role of participating in the control group, i.e., they complete baseline and "posttest" assessments per control group individuals in a classic two-group, pretest-posttest design. Here is the design using the Campbell and Stanley notation:

Treatment Group:	O_1 X O_2
Control Group:	O_1 O_2 O_3 X O_4

The classic analysis of this design is to work with the O_1 and O_2 data for the two groups, analyzing it as if it is a classic two-group pretest-posttest design. However, certain questions also can be addressed by pooling data from O_1 and O_2 for the treatment group with data from O_3 and O_4 for the control group. Wait-list designs typically use passive rather than active control groups at O_1 and O_2 and are challenging if one seeks to assess treatment versus control group effects at extended follow-ups because the wait list group must be delayed that much longer before beginning treatment. These designs are not ideal, but we often have little choice but to use them because of ethical considerations of otherwise denying people treatment.

A related design is a **cross-over design** in which more than one treatment is administered to the same individuals but in different sequences depending on the group one is assigned to. For example, for the two treatments, X_A and X_B, the design might be:

Group 1:	O_1 X_A O_2 O_3 X_B O_4
Group 2:	O_1 X_B O_2 O_3 X_A O_4

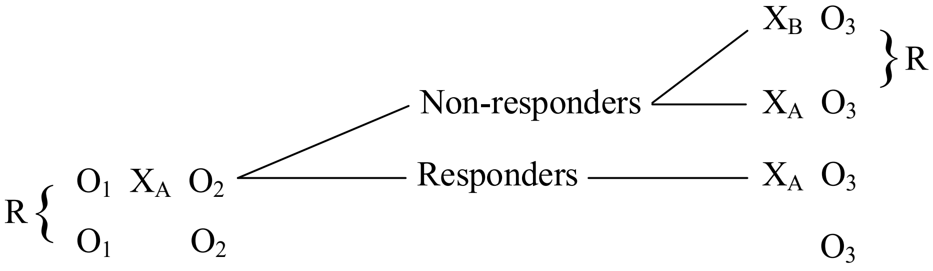
This type of design might be used for evaluating two different headache medications whose effects are short-lived, such as for relief of acute pain during a headache. Medication A is given first for those in Group 1 and then medication B some 6 months later. The reverse order is used for Group 2. The assumption is that the effects of the prior treatment do not carry over to the later treatment. A variant adds a control group, as follows:

Group 1:	O_1 X_A O_2 O_3 X_B O_4
Group 2:	O_1 X_B O_2 O_3 X_A O_4
Group 3:	O_1 O_2 O_3 O_4

For more details about cross over designs, see Jones and Kenward (2014) and Lui (2016).

Trial Design 4: Adaptive Designs

A fourth type of trial design is an ***adaptive design***. Definitions of adaptive designs vary, but in the current book, I define them as designs that (a) randomly assign individuals to a treatment or a control condition, (b) identify individuals in the treatment condition who at a pre-determined point during treatment are not responding well to the treatment, and then (c) randomly assigning these non-responders to either a new treatment regimen or to a condition that continues the original treatment. Usually the control group is assessed throughout all phases of the design, although this varies. Here is the design schematic:



where R indicates random assignment. Comparing the four group means at the final posttest (O_3) is scientifically informative but there are special analytic considerations that must be taken into account given that assignment to the responder versus non-responder groups is non-random.

In sum, there are a host of RCT and RET trial designs that you will encounter. The most common one is the two-group pretest-posttest design or some variant of it. These designs capture the bulk of my attention in this book, although I also consider randomized cluster designs as well.

POPULATIONS FOR RANDOMIZED TRIALS

Almost all of the statistical methods used to analyze data in the social sciences assume that we analyze random samples from a broader population. The analysis of randomized trials is no exception. Do not confuse the practice of selecting for study a random sample from a population with that of randomly assigning individuals to different treatment conditions. They are distinct processes and have different purposes. I discuss each of them in this section, but I first focus on the concept of selecting a random sample from a population, not random assignment to conditions.

Two Ways of "Selecting" Random Samples from a Population

In social science research, the traditional way of thinking about populations and samples involves two steps. First, we define the population of individuals we want to make statements about. Second, we enact procedures that select a random (or approximately random) sample from that population. We then analyze the data in ways that take into account the sampling error that inevitably results when estimating population parameters from the random sample. In idealized random sampling, one has a numbered list of all members of the population and then uses a random number table to select people to invite to be in the study. Given that everyone contacted accepts the invitation, the result is a random sample from the population. This idealized process is often unrealistic and variants of it have been developed that lead to reasonably good approximations of random samples (e.g., area sampling). For details see Blair and Blair (2014).

It is possible to turn this two-step logic on its head by reversing the process. Suppose I conduct a study on a group of recruited individuals, such as patients in a clinic, and then I declare that the group represents a random sample from some population. The task is to then specify the population the sample can be construed as a random sample from. Note that we are still dealing with a population and a random sample from that population. It is just that we are using the sample to drive specification of the population rather than vice versa. I call the former strategy a **population-then-sample** approach and the second strategy a **sample-then-population** approach. Using the latter, I might select people into my study based on responses to flyers I put in a clinic and advertising on the radio or on public transportation. People who respond to these solicitations and who ultimately agree to be in the study are my sample. I then construe them as a random sample from a broader population, which allows me to apply the statistical methods we learn in statistics classes that assume random sampling. The question then becomes just who the broader population is?

The case we make about who the broader population is depends on the procedures we used to recruit the sample, the variables we are studying and the contexts in which those variables are studied. Many scientific researchers avoid thinking about such generalizability by putting an obligatory sentence in their Discussion sections about not generalizing results beyond the "study population" without ever specifying who the "study population" is. I personally think we should demand a more thoughtful consideration of who the population is that the study sample represents.

When we are hired to conduct a program evaluation, the focus usually is on a population that represents the clientele of the clinic, the school, or the organization in which the program is administered. We use recruitment procedures that we think will approximate a random sample from the population that the program serves. This represents a population-then-sample orientation. However, if the recruitment procedures we use result in a non-random sample from the population served, we still analyze the data using statistical methods that assume random sampling. We just shift our mentality to the sample-then-population logic and generalize our conclusions to the population the sample supposedly represents, what I call a **meta-population**.

In actuality, the initial population we seek to generalize to in the population-then-sample framework is not just the population of clientele who are currently served by the organization. Rather, we seek to generalize as well to future clients who will use the services of the organization in the coming years but who have not yet done so. In scientific research, if we seek to establish general laws of human behavior, our populations are even more broad, representing people in the past, people who are currently living, and people from future generations who may not even yet been born. In this sense, our populations are often hypothetical in nature, i.e., they are meta-populations.

When Sampling Bias Does not Matter

Suppose I want to characterize the attitudes of people in the United States about legalizing marijuana. It would be folly for me to conduct a study on college students in a large Northeastern university, assess how favorable the students feel toward legalizing marijuana, and then claim that their opinions can be construed as if the students are a random sample of the general United States population. On the other hand, suppose I want to characterize the effect of smoking marijuana on brain physiology and I again conduct my study on college students in a large Northeastern university. I find that smoking marijuana impacts anandamide molecules in the hippocampus. I might argue that *for these particular variables* and *for this particular question*, the college students essentially function as a random sample of people in the United States and the results can be generalized accordingly. This latter example drives home an important point, namely that biased samples in a technical sense can yield unbiased population estimates. Suppose a population has 50% males and 50% females. A researcher is interested in estimating the divorce rate in the population and, unbeknownst to the investigator, the true overall divorce rate is 40%. Suppose that the divorce rate for males is 40% and it also is 40% for females. Stated another way, biological sex is unrelated to divorce rates. Suppose I conduct a study where my sampling frame, for whatever reason, oversamples males relative to females by a 3 to 1 margin. If my goal is to estimate the overall population divorce rate, the sex bias in my sample is irrelevant; I would get the same result if I used a sampling frame that included equal numbers of males and females because sex is unrelated to the parameter being estimated. The sample bias in biological sex is moot.

Bias only matters for variables that are relevant to estimation of the parameter in question, not variables that are irrelevant to that parameter.

Some argue that when the focus is on basic biological or mental processes, it is not unreasonable to assume one's sample can be construed as a random sample from people in general *for the target variables in question and their interrelationships*. For a study that addresses how psoriasis is impacted by a new drug, a sample of volunteers from a psoriasis clinic in Buffalo, New York, the argument goes, probably can be construed as, functionally, a random sample from a large portion of adults in the United States. Or can it? The onus is on the researcher to make a case for the population the sample represents.

Populations in Randomized Trials

The specification of a population based on a sample is more complicated for the case of randomized trials. Before describing how, let me first reiterate some key points about random assignment to conditions. In a typical trial, randomizing people to conditions means that each person has an equal chance of being in the treatment or control condition. A corollary of randomization is that it is equally likely that a person with any given attribute, A, will be in the treatment condition as in the control condition. Randomization to conditions tends to produce samples that are comparable at baseline on all known as well as unknown outcome determinants, including the baseline outcome itself. In this sense, randomization is said to control for pre-treatment confounds when evaluating the effects of a treatment relative to a comparison condition.

Most randomized trials use convenience samples and adopt a sample-then-population approach to specifying the study population. The population is further defined as representing two sub-populations that are equivalent in all respects except one, namely whether population members have experienced the treatment or whether they have experienced the comparator. The two hypothetical populations, by virtue of random assignment, are thought to have equal baseline means on the outcome but their posttest means can differ because of the effect of the intervention on individuals in the "treatment" population. Not only is the baseline mean presumed equal in the two populations, this also is the case for any baseline variable that might impact the outcome. This conceptualization of the populations is important to keep in mind when addressing imbalance due to randomization, a topic I now address.

THE CONCEPT OF IMBALANCE

When analyzing data for a randomized trial, researchers typically calculate for each condition the sample means of the outcome at baseline, $\overline{Y}_0$, and the sample means at the postest, $\overline{Y}_1$. These statistics represent estimates of the baseline and posttest means for the two sub-populations noted above. There, of course, will be sampling error in the means relative to the true population means and this also will be true for any determinant of Y that I measure, either at baseline or at the posttest. When treatment and control baseline sample means on some variable are not exactly equal, the data are said to be *imbalanced*. Even though the population means for baseline variables are, in theory, balanced (because of random assignment), the sample baseline means may not be balanced because of sampling error. The theory of random assignment holds that across many replications of a trial, the biasing effects of these imbalances cancel when sample estimates are averaged across replications. However, it does not change the fact that when working with data for just one trial, sample imbalance not only can result but it is fully expected given the presence of sampling error.

Imbalance as an Indicator of Compromised Randomization

If large amounts of sample imbalance are observed in a given study, then this might lead an investigator to question if random assignment to conditions truly occurred. As discussed earlier, sometimes random assignment is compromised, such as when staff do not follow randomization protocols. If randomization is faithfully executed, however, then with a large enough sample size, we do not expect to observe much imbalance in the sample data because there should be little sampling error. If we do observe large imbalance, then this might suggest that the two referent populations are not equal on baseline variables because randomization was compromised. To explore this possibility, some researchers perform tests of statistical significance on the baseline variables contrasting the treatment and control conditions with the idea that such tests can determine if random assignment was compromised. This practice is controversial, with most methodologists recommending against it (e.g., Bland & Altman, 2011; Senn, 1994).

The case against conducting between-group baseline significance tests if we do *not* think randomization is in doubt has been aptly stated by Altman (1985):

> *performing a significance test to compare baseline variables is to assess the probability of something having occurred by chance when we know that it did occur by chance given proper implementation of randomization.* Altman (1985, p. 126)

If, however, we believe randomization may have been compromised, the use of significance tests for evaluating this possibility is not illogical; it is just that many methodologists consider it to be a weak strategy for determining compromised randomization. Objections include (a) the fact that such tests usually examine many baseline variables with the consequent problem of inflated Type I error rates due to multiplicity, (b) the statistical power of such tests might be low causing us to miss meaningful imbalance, (c) if the sample size is large, then the tests might detect levels of imbalance that are trivial, and (d) there is a lack of clear standards surrounding what constitutes a sufficient between-group difference on variables to question randomization (de Boer, Waterlander, Kuijper, Steenhuis, & Twisk, 2015). To be sure, all methodologists agree that researchers should be sensitive to the possibility of compromised random assignment. However, significance testing of baseline differences is seen as a weak approach to gaining perspectives on the matter.

Imbalance and Small Sample Sizes

Imbalance between groups for sample data will tend to be less the larger the per group sample size. This is because there is smaller amounts of sampling error with larger N. With small N, however, it is possible that non-trivial sample imbalance will occur, even when there is no imbalance in the populations. Hsu (1989) estimated the probability that "non-trivial" sample imbalance would occur as a function of sample size for a randomized two-group design. Non-trivial imbalance was defined for a dichotomous variable, X (scored 0 and 1), when either the treatment or the control group had twice as many people with a score of 1 on X than the other group after random assignment. In addition to varying sample size, Hsu calculated this probability for differing numbers of baseline outcome determinants (also called **prognostic variables**). For example, if there were three baseline variables that impact the outcome (meaning imbalance on them is potentially consequential), Hsu calculated the probability that at least one of them would show "non-trivial" false imbalance. Table 4.2 presents his results. As an example for reading the table, for a total sample size of 18 (9 per group), the probability that at least one prognostic variable given two such variables would show "non-trivial" false imbalance is 0.574. It can be seen that for N = 64 (32 per group) or higher, the chances of non-trivial false imbalance was quite low. This suggests that of your N is about 65 per group or larger, sample imbalance probably is not an issue.

Strube (1991, 2015) argued that a more accurate picture of the consequences of false imbalance requires not only documentation of how often it occurs as a function of sample size, but also how it affects posttest significance tests of group differences on the outcome. This is important because the tendency for small-sample induced false imbalance to create a false treatment effect might be offset by the reduced power to detect that false treatment effect due to the small N. Strube found that false imbalance only induced unacceptable Type I error rates when the effect size of the prognosticator was large (Cohen's d > 2.0 or more) for sample sizes less than 15 or so per group. He concluded that "although the likelihood of nonequivalence may be quite high for small samples…, the likelihood is quite low that the nonequivalence will produce erroneous inferences about treatment efficacy." (p. 349).

Table 4. 2: Probability of False Imbalance

	Number of Prognostic Variables		
Total N	1	2	3
8	0.486	0.736	0.864
12	0.567	0.813	0.919
18	0.347	0.574	0.721
24	0.220	0.392	0.526
32	0.076	0.146	0.210
40	0.026	0.050	0.075
64	0.005	0.011	0.016
80	0.003	0.007	0.010
100	0.001	0001	0.002

Addressing Imbalance

Given the inevitable presence of imbalance in sample data (but assuming there is no imbalance in the populations), the question becomes what to do about it, if anything. Here are some guidelines. First, sample imbalance on variables that do not impact the outcome (or the mediators) is irrelevant and can safely be ignored. We are only concerned with imbalance on variables that matter. For example, if the treatment and control conditions differ on shoe size, who cares? Second, for prognosticators of Y *that show non-trivial imbalance*, most methodologists recommend controlling for them during data analysis (Altman, 1985; de Boer et al., 2015; Greenland, Robins, & Pearl, 1999; Senn, 1994, 2013). This recommendation might seem surprising in light of the results of Hsu and Strube, but the idea is that doing so can only improve our estimates of program effect sizes, so why not adjust for them. There are, of course, qualifications to this suggestion. For example, if the measures of the covariates that show non-trivial sample imbalance have low reliability, then covarying them out may make matters worse, not better. If there are outliers on the covariate that distort its relationship to the outcome, then including it in the model can make matters worse unless you deal with those outliers. Given this and other modeling complications, you will want to be judicious about what covariates you control to deal with sample imbalance (Senn, 2013).

In practice, I routinely check for baseline imbalance on variables I have measured before embarking on data analysis, although I do not rely on tests of statistical significance to do so. I focus instead on effect size indices of baseline differences. If I find a potentially troublesome effect size for imbalance for a given variable, I ask myself if the variable is a likely determinant of either my mediators or outcomes. If the answer is no, then the imbalance is irrelevant. If the answer is yes, I consider controlling it during data analysis by including it as a covariate in the analysis. I usually prioritize my list of covariates in terms of importance and then control as many as is practically reasonable.

RANDOMIZATION STRATEGIES

There are many methods for randomly assigning people to treatment versus control conditions. In this section, I describe four approaches, fixed allocation, block randomization, stratified randomization, and adaptive randomization.

Fixed Allocation Random Assignment

Fixed allocation methods assign people to treatment or control conditions based on a pre-specified probability, usually equal (1:1). Some methodologists suggest the use of 2:1 allocation ratios, with the N for the intervention group being twice the size of the control group. This allows researchers to address within treatment analyses that might not otherwise be possible due to small N. Such allocations must be used with caution because the statistical power of a two-group contrast is primarily driven by the smaller sample size of the two groups; a 2:1 allocation can thus reduce power relative to a 1:1 scheme, sometimes substantially so.

For simple randomization, one can use computer software to generate random numbers that determine which cases to assign to which condition. Table 4.3 presents example output from the programs available on this book's website using a sample size of 20 per group. If the list of participants is available before random allocation, then the cases are numbered from 1 to 40 and assigned per the table. If individuals are instead recruited and assigned to a condition sequentially in the order they appear at a clinic or organization, the numbers in the table refer to the sequence number of the participant (1 = first recruited participant, 2 = second recruited participant, and so on through 40 = last recruited participant), with condition assignment dictated accordingly. For example, the second person who is recruited into the study will be assigned to Condition 1 using Table 4.3. but the first person recruited into the study will be assigned to Condition 2.

Table 4. 3: Simple Random Assignment

```
ASSIGN THE FOLLOWING CASES TO CONDITION 1  (N = 20)

2    10    16     24
5    11    19     31
6    12    20     32
7    13    21     36
9    15    23     37

ASSIGN THE FOLLOWING CASES TO CONDITION 2  (N = 20)

1    17    27     34
3    18    28     35
4    22    29     38
8    25    30     39
14   26    33     49
```

A variant of this approach is to flip a coin for each participant as they are recruited into the study, with heads indicating assignment to the treatment condition and tails indicating assignment to the control condition. An advantage of this strategy is that the researcher or staff member making the assignment has no idea which condition a person will be assigned to until that very second. Chalmers, Celano, Sacks and Smith (1983) reviewed reports of over 100 clinical trials. For studies where the researcher or staff member was blind to the condition to which the person would be assigned, 14% of the studies observed imbalance on at least one baseline variable. For studies where the researcher or staff member was not blind to the treatment condition a person would be assigned to, the corresponding imbalance rate was 26%.[1] The pre-generated list per Table 4.3 can be used in a blinded way if staff call a centralized data coordination location and are told over the phone just prior to allocation the condition the person is to be assigned to. Or, staff can learn of condition assignment at the time of allocation from a project web page they consult that monitors allocation vis-à-vis Table 4.3.

Some researchers believe an alternating assignment of sequentially recruited participants to the treatment or control condition constitutes random assignment (e.g., T-C-T-C-T-C-…., where T is the treatment condition and C is the control condition). Such a listing has no random component to it, so it does not fulfill random assignment requirements. In random assignment, the group to which the next case is to be assigned should be unpredictable. This is not the case with an alternating sequence.

[1] These results must be taken with a grain of salt because the imbalance was defined on the basis of significance tests.

Block Random Assignment

Another randomization strategy is called **block randomization**. This method is used if one wants to avoid condition imbalance in the number of participants in a condition across time, given sequential recruitment of people into the trial. It assures that at certain points in time, the number of people in each condition will be equal. As such, block randomization protects against temporal trends during enrollment if one suspects that meaningful external conditions might be changing across the time of recruitment. For example, if one is recruiting 7th grade adolescents into a study of math achievement over the course of an academic year, students recruited early in the academic year will likely have less math achievement than students recruited later in the academic year, due to naturally occurring brain development and the fact that students are learning math in their classes over the course of the year. Simple randomization, properly executed, generally will work in these cases, but if one wants to assure equal numbers of participants in the two conditions as time unfolds, block randomization can be used.

In block randomization, suppose one wants to ensure that after every fourth person randomized, the number of people in the two conditions is equal. The block size is defined as 4, with each block containing two treatment assignments and two control assignments. For two conditions and block sizes of 4, there are 8 possible block types (e.g., T-C-T-C; T-T-C-C; C-T-C-T, and so on). One of these block types is randomly selected and people are assigned in accord with the ordering within the selected block. The process is repeated after the first block is complete. A disadvantage of blocked randomization is that most statistical methods for data analysis assume simple random sampling. With block randomization, special analytic adjustments are required. For details, see Calinski and Kageyama (2003) and Rosenberger and Lachin (2015).

Stratified Random Assignment

Stratified randomization is a random assignment strategy designed to reduce sample imbalance. It requires having information about relevant prognosticators for each participant before random assignment. Continuous prognosticators usually are divided into strata (e.g., if the prognosticator is age, three age groups might be defined). The strata that a recruit is in is determined and then assignment to the treatment or control condition is randomly determined within that strata. Strata can be defined for more than one prognosticator, with each prognosticator being treated as a "factor." The factorial combination of prognosticators then defines the stratification design. Like block randomization, stratified randomization requires specialized statistical methods; see Rosenberger and Lachin (2015).

Adaptive Random Assignment

Adaptive randomization alters the allocation ratios as a trial progresses. There are many types of adaptive randomization. **Covariate adaptive randomization** sequentially assigns participants to treatment conditions in ways that consider prior participant assignments and baseline prognosticators. The first *n* of N participants (e.g., 20% of them) are assigned to condition using simple randomization. At that point, each subsequent participant is assigned to a condition based on his or her scores on the targeted prognosticators. Specifically, an aggregate index of prognosticator values (e.g., mean scores across prognosticators) are calculated for each condition based on those individuals already assigned to the groups. The new participant is then assigned to the condition that will cause the aggregate index in the two conditions to be closer in value, thereby reducing imbalance. Different types of adaptive randomization are defined by variants of methods for aggregation and for the decision rules to assign participants to conditions based on those aggregates. For details, see Rosenberger and Lachin (2015).

In sum, there are multiple randomization strategies, with the most common strategy being simple randomization. I concentrate on it in this book. The other methods have evolved largely in the context of outcome-only randomized trials rather than RETs. Extending them to account for multiple mediators as well as multiple outcomes in a single RET can be difficult in applied settings.

PHASES OF A RANDOMIZED TRIAL

A well-known way of describing the phases of a randomized trial is the approach used by the Consolidated Standards of Reporting Trials (CONSORT) organization. There are four phases, (1) enrollment, (2) allocation, (3) follow-up, and (4) data analysis. CONSORT (Moher, Hopewell, Schulz, Montori, Gøtzsche, Devereaux, et al., 2010) developed a flow-chart to document participant flow through these phases (see Figure 3.1). Most scientific journals require researchers to report the chart. The first phase is enrollment, in which individuals are screened for trial eligibility using explicit inclusion and exclusion criteria. Researchers document the number of people screened, the number who were excluded, and the reasons for exclusion. This process defines the number of people who are randomized.

The second phase is allocation to the treatment conditions, where one of the conditions usually is the treatment group and the other is the control group. The CONSORT chart in Figure 3.1 shows the label "treatment" for each group because CONSORT uses the term "treatment" in its most generic sense. The chart can be edited to add more than two groups and it can explicitly label one of them the "control" or "no treatment" group, as appropriate. Within each group, you indicate the number of people who received the allocated intervention and the number who did not and why.

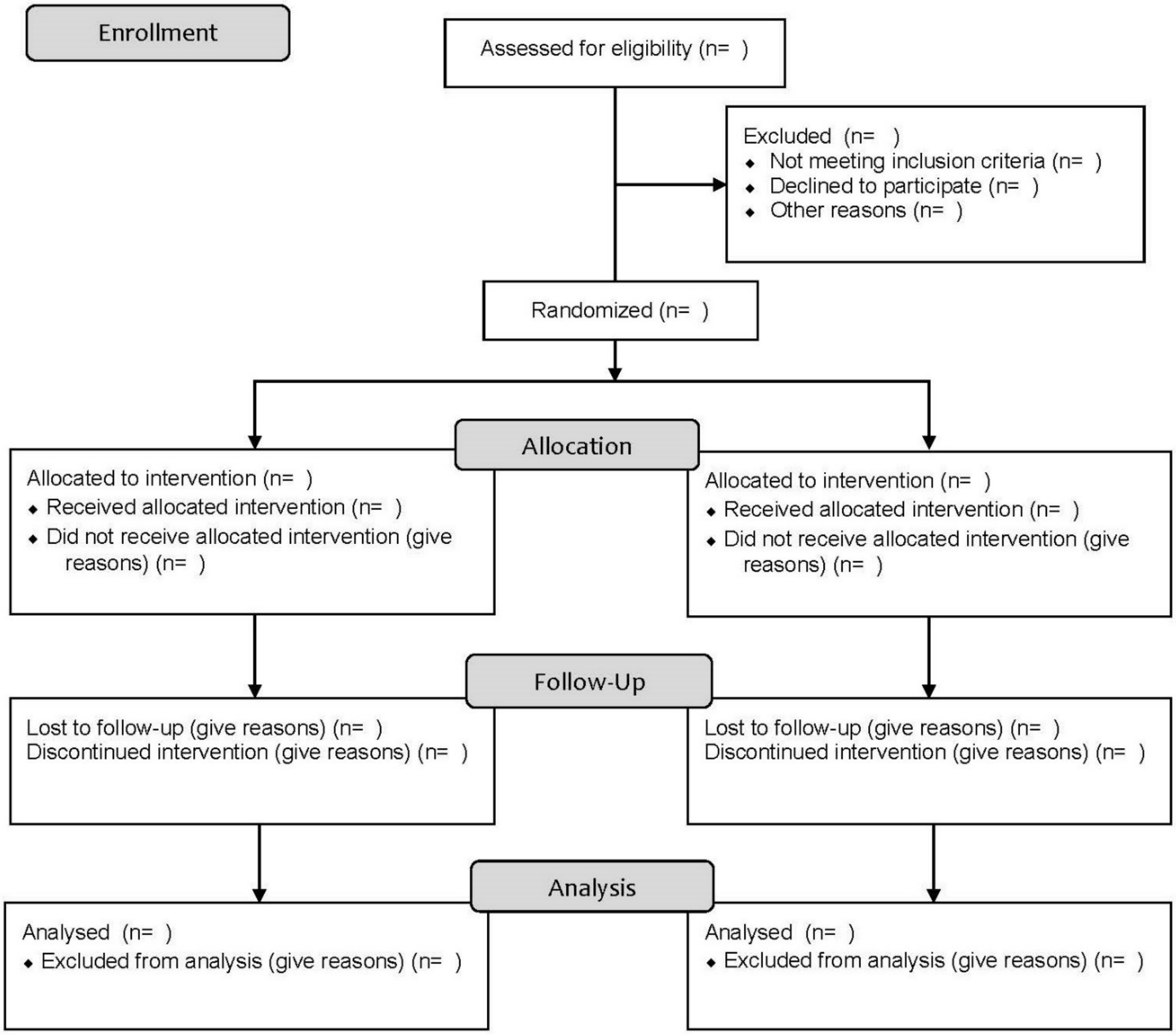

Enrollment
Assessed for eligibility (n=)
Excluded (n=)
◆ Not meeting inclusion criteria (n=)
◆ Declined to participate (n=)
◆ Other reasons (n=)
Randomized (n=)
Allocation
Allocated to intervention (n=)
◆ Received allocated intervention (n=)
◆ Did not receive allocated intervention (give reasons) (n=)
Allocated to intervention (n=)
◆ Received allocated intervention (n=)
◆ Did not receive allocated intervention (give reasons) (n=)
Follow-Up
Lost to follow-up (give reasons) (n=)
Discontinued intervention (give reasons) (n=)
Lost to follow-up (give reasons) (n=)
Discontinued intervention (give reasons) (n=)
Analysis
Analysed (n=)
◆ Excluded from analysis (give reasons) (n=)
Analysed (n=)
◆ Excluded from analysis (give reasons) (n=)

FIGURE 3.1. CONSORT flowchart

The third phase is the follow-up phase, with the follow-up period starting after treatment initiation. You report the number of participants who discontinued treatment and the reasons why this occurred, as well as the number of participants who were "lost to follow-up," i.e., the number of people who could not be contacted for post intervention assessments. The final phase is the data analysis phase. Here, you indicate the number of participants who were excluded from the analysis and why.

The chart is oversimplified for many trials. Researchers are expected to amend the chart as needed and to provide details of the amendments in the text of the report. The spirit of the chart is to provide readers with a clear record of participant flow through the trial phases and to provide an accounting of reasons why there is fall-off in the flow at the different phases.

You also will encounter the concept of "trial phases" in a completely different context, namely with reference to a sequence of studies that move from the initial phases of intervention testing to widespread intervention implementation. The number and characterization of these phases varies, but usually there are four of them, traditionally framed in medical terms. A **Phase I** trial involves the testing of a new drug or treatment on a small group of people to determine safe dosage levels, toxicity, and major side effects. A **Phase II** trial evaluates the efficacy of the drug or treatment on a small-scale basis and builds a case for treatment potential. A **Phase III** trial is administered to larger groups of people to confirm efficacy, to evaluate effectiveness of the treatment, to compare it to other commonly used treatments, and/or to gain further insights on safety. A **Phase IV** trial further evaluates the drug or treatment in the general population after it has been licensed and marketed. RETs can be pursued during all phases.

DEMAND CHARACTERISTICS AND BLINDING

In psychology, the concept of a demand characteristic is a well-known experimental artifact that researchers routinely address. A **demand characteristic** is a cue that makes study participants aware of what an experimenter might expect to find in the study or how study participants are expected to behave. Participants then purposely alter their behavior and/or responses to questionnaires to match these perceived expectations.

In a randomized trial, by virtue of informed consent, individuals know they are participating in an evaluation of a treatment/program. Many participants may infer that the "new" program will be effective and/or that the investigator wants the program to succeed. They therefore might positively bias their characterizations of the program and the self-report measures designed to assess program outcomes. In a single group pretest-posttest design, such demand characteristics can make a treatment appear more effective than it is. For a randomized trial, demand characteristics are likely to operate in both the treatment and control conditions (if an active control is used) or if individuals are unaware of which condition they are in. Given this, between-group comparisons of outcome distributions help to control for demand characteristics. The possible operation of demand characteristics is yet another reason we prefer between-group treatment versus control means to within-group analyses of pretest to posttest mean change.

A **double-blind** study is one in which neither the participants nor the experimenters know who is receiving a particular treatment condition. It is a standard that is sought under the assumption that it reduces not only the impact of demand characteristics on the part of participants but also other forms of bias on the part of the staff that knowledge of the treatment condition might produce. Although a double-blind trial is desirable, it does not alter the fact that participants know they are in a randomized trial and that an intervention/treatment is being evaluated. This is why treatment versus control comparisons should be the standard by which a treatment is evaluated.[2]

There are steps one can take to increase honest responding to help reduce bias due to demand characteristics. These include: (1) use of a self-administered as opposed to a face-to-face reporting format so that participants do not have to report sensitive behaviors directly to an interviewer; (2) ensure response confidentiality and assure participants that identifying information will not be associated in any way with their data; (3) stress that results in research reports will only be reported for groups of people, not for individuals; (4) emphasize to participants that the quality and ability of the research to "make a difference" depends on people giving honest answers; (5) instruct respondents not to answer a question if they are not going to be truthful (and then use modern analytic methods to handle the missing data); and (6) obtain a measure of participant social desirable response tendencies to use as a potential statistical covariate during modeling.

TREATMENT INTEGRITY

Treatment integrity refers to the degree to which an intervention or program is implemented as intended. In randomized trials assessments of treatment integrity are considered to be good methodological practice, although researchers often fail to do so (Perepletchikova, Treat, & Kazdin, 2007). For efficacy trials, correctives are often introduced if treatment integrity falls below a pre-specified threshold during monitoring. For effectiveness trials, treatment infidelity is a natural part of program evaluation, so correctives are not introduced. Instead, treatment integrity is stressed when the intervention is first introduced to a clinic or organization. There is a large literature in implementation science that addresses how best to maximize treatment integrity in clinics or organizations. For introductions to this literature, see Perepletchikova (2011) and Brownson, Colditz, & Proctor (2017).

[2] Wait-list control designs usually are inadequate for controlling demand characteristics because control individuals know they have not yet started treatment

SELECTION EFFECTS AND GENERALIZABILITY

Earlier I discussed selection effects as a threat to the internal validity of an RET. There also are forms of selection that threaten the generalizability of one's results, either in a population-then-sample or a sample-then-population sense. For example, people might self-select into or out of my study based on a positive or negative response to flyers I put in a clinic and advertising on the radio or on public transportation inviting them to contact my office if interested in study participation. Or, I might approach a person to participate in my study but the person declines to do so. Randomization to conditions typically occurs after people have self-selected into the study, so the internal validity of the study is unaffected by such selection. However, the self-selection dynamics *can* impact who the internally valid results apply to. If I develop a program to teach parents how best to talk with their adolescent children about sex but only motivated parents who are prone to talk with their adolescent children about sex agree to be in my study, then perhaps my study results do not apply to parents who are reluctant to talk with their children about sex, thereby limiting result generalizability.

When I conduct an RET, I try to document if meaningful selection bias has occurred. I often use two strategies, neither of which is perfect. First, for a subsample of decliners, I ask them if they can complete a very brief questionnaire that takes only a few minutes and that I will monetarily compensate them for their time to complete. I explain that it is scientifically important to document general characteristics of people who can't or choose not to be in my study. The questionnaire or interview I give them has 5 to 10 carefully selected questions that I also ask of people who agree to be in the study, perhaps as part of the regular study protocol. I then conduct comparisons between decliners and participators for self-selection bias on the questions to determine if self-selection bias on those questions has occurred. I typically find that almost all decliners are willing to help out in this way.

If the above strategy is not feasible, I use a second method, or, ideally, I use both methods. Suppose my study is conducted in a clinic or community for which I have aggregate level statistics from another source, such as the percent of different ethnic groups, the prevalence of risk factors, and so on. If my sample is not contaminated by self-selection bias, then it should mirror these aggregate level statistics. Given such mapping, I have more confidence that self-selection into my study has not intruded on result generalizability. Of course, this strategy is limited by the nature of the aggregate level data I have available and it can miss self-selection bias that has occurred but for which I do not have aggregate data. However, if I can use such data to strengthen my case for the absence of meaningful self-selection bias, I take the opportunity to do so.

REGISTERING CLINICAL TRIALS

For many clinical trials, it is common practice to register the trial with the federal government on a site called www.clinicaltrials.gov. Registered trials are those that use humans to assess biomedical and/or health outcomes and that conform to applicable ethics review regulations. Table 4.4 presents an example from the ClinicalTrials.gov website of the information typically needed to register a trial (in addition to IRB approval).

Table 4.4: Registration Information

Overview

Summary: The purpose of this study is to evaluate, subjectively and objectively, whether playing music during procedures for treatment of chronic lower back pain has an effect on patients' anxiety and pain. Our hypothesis is that playing music will result in reduced patient reported anxiety and pain scores and less variation from baseline of vital signs versus patients in the control group without music therapy. This is a pilot study.

Condition or disease: Chronic pain, anxiety

Intervention/treatment: Music Therapy, No music

Phase: Not Applicable

Study Type: Interventional (Clinical Trial)

Estimated Enrollment: 30 participants

Allocation: 1:1 Randomized

Intervention Model: Single Group Assignment

Masking: None (Open Label)

Primary Purpose: Treatment

Official Title: Effect of Music on Pain and Anxiety in Chronic Pain Patients
undergoing Lumbar Intervention
Procedures: A Pilot Study

Estimated Study Start Date: June 15, 2021

Estimated Primary Completion Date: December 1, 2021

Estimated Study Completion Date: December 1, 2021

Arms and Interventions

Arm: Intervention/treatment

Active Comparator: No Music

No music will be played during the subject's standard of care lumbar spinal interventional procedure (including epidural steroid injections, facet injections, medial branch blocks).

Experimental: Music Therapy

Music of the subject's preferred genre will be played during the subject's standard of care lumbar spinal interventional procedure (including epidural steroid injections, facet injections, medial branch blocks).

Outcome Measures

Primary Outcome Measures:

Measure 1: Pre-procedure STAI Score
[Time Frame: Within 30 minutes prior to the subject's interventional procedure]
A measure of anxiety in a person

Measure 2: Post-procedure STAI Score [Time Frame: Within 30 minutes following to the subject's interventional procedure]
A measure of anxiety in a person

Secondary Outcome Measures:

Measure 1: Pre-procedural Visual Analog Score (VAS) for pain
[Time Frame: Within 30 minutes prior to the subject's interventional procedure]
Visual Analog Score for pain

Measure 2: Post procedural VAS Score
[Time Frame: Within 30 minutes following the subject's interventional procedure]
Visual Analog Score for pain

Eligibility Criteria

Ages Eligible for Study: 18 Years and older (Adult, Older Adult)

Sexes Eligible for Study: All

Accepts Healthy Volunteers: No

Patients undergoing standard of care lumbar spinal interventional procedures including epidural steroid injections, facet injections, medial branch blocks

Exclusion Criteria:

Patients who cannot consent for themselves, including cognitively impaired patients.
Non-English speaking patients
Patients taking beta blocker medication
Patients that have a pacer and have a set rate
Patients with self-reported hearing problems or with hearing aids

Many researchers also publish a detailed study protocol in journals such as *Trials.* These publications allow researchers to present detailed information about conceptual logic models for a trial, general trial methodology, measurement approaches, and statistical analysis plans, all of which can then be referenced in later publications. The spirit of journals like *Trials* also is to promote transparency.

THE CONSORT CHECKLIST

As a final note, many journals require researchers to complete a CONSORT checklist for randomized trials. I reproduce an abridged version in Table 4.5 because it provides a useful summary of issues to attend to in trial design. In addition to the CONSORT statement, the Cochrane Group (2019) has published a checklist for assessing "risk of bias" of randomized trials. Although there are many methodological points in that list you will want to attend to, I personally find the algorithm for combining them into an overall risk score to be somewhat dubious.

Table 4.5: Abridged CONSORT Checklist

		Reported on Page Number
Title and abstract		
1a	Identification as a randomised trial in the title	________
1b	Structured summary of trial design, methods, results, and Conclusions (for specific guidance see CONSORT for abstracts)	________
Introduction		
Background and objectives		
2a	Scientific background and explanation of rationale	________
2b	Specific objectives or hypotheses	________

Reported on
Page Number

Methods

Trial design

3a Description of trial design (such as parallel, factorial) including allocation ratio ________

3b Important changes to methods after trial commencement (such as eligibility criteria), with reasons ________

Participants

4a Eligibility criteria for participants ________

4b Settings and locations where the data were collected ________

Interventions

5 The interventions for each group with sufficient details to allow replication, including how and when they were actually administered ________

Outcomes

6a Completely defined pre-specified primary and secondary outcome measures, including how and when they were assessed ________

6b Any changes to trial outcomes after the trial commenced, with reasons ________

Sample size

7a How sample size was determined ________

7b When applicable, explanation of any interim analyses and stopping guidelines ________

Randomisation:

Sequence generation

8a Method used to generate the random allocation sequence ________

8b Type of randomisation; details of any restriction (such as blocking and block size) ________

Allocation concealment mechanism

9 Mechanism used to implement the random allocation sequence (such as sequentially numbered containers), describing any steps taken to conceal the sequence until interventions were assigned ________

Implementation

10 Who generated the random allocation sequence, who enrolled participants, and who assigned participants to interventions ________

Reported on Page Number

Blinding

11a If done, who was blinded after assignment to interventions (for example, participants, care providers, those assessing outcomes) and how ________

11b If relevant, description of the similarity of interventions ________

Statistical methods

12a Statistical methods used to compare groups for primary and secondary outcomes ________

12b Methods for additional analyses, such as subgroup analyses and adjusted analyses ________

Results

Participant flow (a diagram is strongly recommended)

13a For each group, the numbers of participants who were randomly assigned, received intended treatment, and were analysed for the primary outcome ________

13b For each group, losses and exclusions after randomisation, together with reasons ________

Recruitment

14a Dates defining the periods of recruitment and follow-up ________

14b Why the trial ended or was stopped ________

Baseline data

15 A table showing baseline demographic and clinical characteristics for each group ________

Numbers analysed

16 For each group, number of participants (denominator) included in each analysis and whether the analysis was by original assigned groups ________

Outcomes and estimation

17a For each primary and secondary outcome, results for each group, and the effect size and its precision (such as 95% confidence interval) ________

17b For binary outcomes, presentation of both absolute and relative effect sizes is recommended ________

Ancillary analyses

18 Results of any other analyses performed, including subgroup Analyses and adjusted analyses, distinguishing pre-specified from exploratory ________

Reported on Page Number

Harms

19 All important harms or unintended effects in each group (for specific guidance see CONSORT for harms) ________

Discussion

Limitations

20 Trial limitations, addressing sources of potential bias, imprecision, and, if relevant, multiplicity of analyses ________

Generalisability

21 Generalisability (external validity, applicability) of the trial findings ________

Interpretation

22 Interpretation consistent with results, balancing benefits and harms, and considering other relevant evidence ________

Other information

Registration

23 Registration number and name of trial registry ________

Protocol

24 Where the full trial protocol can be accessed, if available ________

Funding

25 Sources of funding and other support (such as supply of drugs), role of funders ________

CONCLUDING COMMENTS

The present chapter has outlined a wide range of methodological issues that researchers need to take into account when designing RETs. RETs need to address core issues of program confounds per Campbell & Stanley (1963), devise effective randomization strategies, address sample imbalance, minimize demand characteristics, and ensure treatment integrity. RETs are, by nature, more complicated than traditional outcome only randomized trials given their joint focus on mediators, moderators and outcomes. Every methodological issue one must think about when designing an outcome-only randomized trial must also be brought to bear on mediators and moderators. Future chapters will identify additional methodological issues that researchers need to consider when designing RETs.