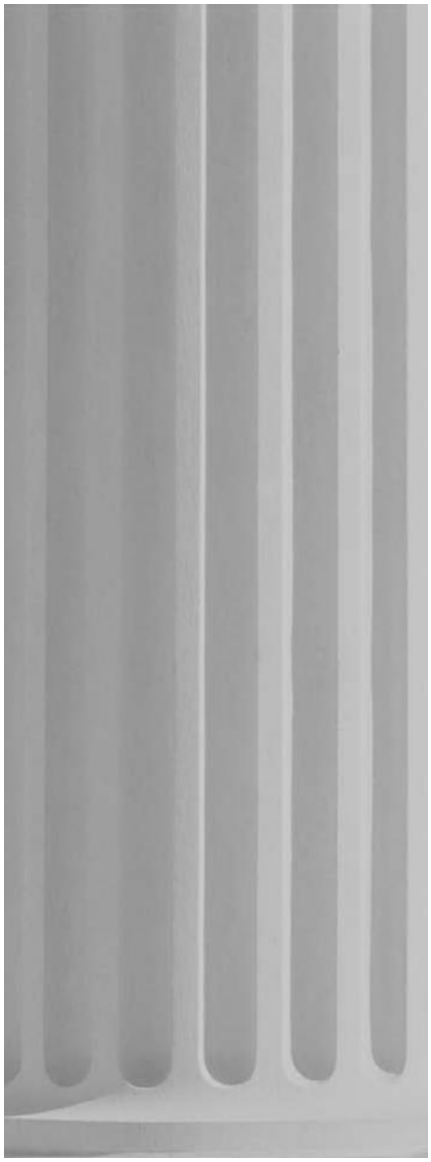
CHAPTER 11

# Advanced Design Topics

## KEY TERMS

causal
control group
covariance
external validity
history threat
internal validity
main effect
mortality threat
NEGD design
pattern matching
random assignment
RD design
RE design
selection bias
threats to internal validity



## OUTLINE



This chapter encourages you to think deeply about social research design. Although I've called this chapter "advanced" design topics, don't let that put you off. Just because they're advanced doesn't mean they're hard to understand. Up to now, I have primarily been talking about specific designs and design types. But all three topics in this chapter talk about issues that cut across the entire research design endeavor. The chapter begins by addressing how you go about designing a research design. I might have put this topic before the chapters on specific designs, but I think that it is better to address this issue after you have a firm foundation in design. The second topic discusses the commonalities across all designs that have a pretest and posttest and a treatment and **control group**. These pre-post, two-group designs are the most common designs used for **causal** assessment. If you can understand the underlying similarities and differences of this type of design, you'll be well on your way to mastering research designs in general. Finally, I conclude with considerations of some of the major hot topics in research design today. Hopefully, this will put you on the cutting edge (and hopefully, you won't get cut!).

**control group**
A group, comparable to the program group, that didn't receive the program.

**causal**
Pertaining to a cause-effect relationship.

# 11-1 Designing Designs for Research[1]

Much contemporary social research is devoted to examining whether a program, treatment, or manipulation causes some outcome or result. For example, you might want to know whether a new educational program causes subsequent achievement score gains, whether a special work release program for prisoners causes lower recidivism rates, whether a novel drug causes a reduction in symptoms, and so on. In Chapter 7, I mentioned that three conditions must be met before you can infer that such a cause-effect relationship exists:

1. *Covariation.* Changes in the presumed cause must be related to changes in the presumed effect. Thus, if you introduce, remove, or change the level of a treatment or program, you should observe some change in the outcome measures.
2. *Temporal precedence.* The presumed cause must occur prior to the presumed effect.
3. *No plausible alternative explanations.* The presumed cause must be the only reasonable explanation for changes in the outcome measures. If other factors could be responsible for changes in the outcome measures, you cannot be confident that the presumed cause-effect relationship is correct.

In most social research, the third condition is the most difficult to meet. Any number of factors other than the treatment or program could cause changes in outcome measures. Chapter 7 lists a number of common plausible alternative

[1]Much of the material for this section is based on Trochim, W., and Land, D. (1982). Designing Designs for Research. *The Researcher*, 1, 1, 1–6.

explanations (or **threats to internal validity**). For example, it may be that some historical event that occurs at the same time that the program or treatment is instituted is responsible for the change in the outcome measures, or changes in record keeping or measurement systems that occur at the same time as the program might be falsely attributed to the program.

**threats to internal validity**
Threats to the the approximate truth of inferences regarding cause-effect or causal inferences.

The typical social science methodology textbook (which this book is not, I daresay) usually presents an array of research designs and the alternative explanations these designs rule out or minimize. This tends to foster a "cookbook" approach to research design—an emphasis on the selection of an available design off the shelf, as it were. While standard designs may sometimes fit real-life situations, top-notch researchers (which I'm sure you aspire to be) learn how to tailor a research design to fit the particular needs of the research context and minimize the relevant threats to validity. Furthermore, even if standard textbook designs are used, an understanding of the logic of design construction in general will improve your comprehension of these standard approaches. In this section, I present an approach to how to design a research design. While this is by no means the only strategy for constructing research designs, it helps clarify some of the basic principles of design logic.

## 11-1a Minimizing Threats to Validity

Before we get to constructing designs themselves, it would help to think about what designs are designed to accomplish. Good research designs minimize the plausible alternative explanations for the hypothesized cause-effect relationship. But research design is not the only way you can rule out threats. Here, I present five ways to minimize threats to validity, one of which is by research design:

1. *By argument.* The most straightforward way to rule out a potential threat to validity is simply to argue that the threat in question is not a reasonable one. Such an argument may be made either *a priori* or *a posteriori.* (That's before the fact or after the fact, for those of you who never studied dead languages.) The former is usually more convincing than the latter. For example, depending on the situation, you might argue that an instrumentation threat is not likely because the same test is used for pretest and posttest measurements and did not involve observers who might improve or change over time. In most cases, ruling out a potential threat to validity by argument alone is weaker than using the other following approaches. As a result, the most plausible threats in a study should not, except in unusual cases, be ruled out by argument alone.
2. *By measurement or observation.* In some cases it is possible to rule out a threat by measuring it and demonstrating that either it does not occur at all or occurs so minimally as to not be a strong alternative explanation for the cause-effect relationship. Consider, for example, a study of the effects of an advertising campaign on subsequent sales of a particular product. In such a study, history (meaning the occurrence of other events than the advertising campaign that might lead to an increased desire to purchase the product) would be a plausible alternative explanation. For example, a change in the local economy, the removal of a competing product from the market, or similar events could cause an increase in product sales. You can attempt to minimize such threats by measuring local economic indicators and the availability and sales of competing products. If there are no changes in these measures coincident with the onset of the advertising campaign, these threats would be considerably minimized. Similarly, if you are studying the effects of special mathematics training on math achievement scores of children, it might be useful to observe everyday classroom behavior to verify that students were not receiving any math training in addition to that provided in the study.
3. *By design.* Here, the major emphasis is on ruling out alternative explanations by adding treatment or control groups, waves of measurement, and the like. I've already covered how you do this in the previous two chapters.

**mortality threat**
A threat to validity that occurs because a significant number of participants drop out.

**main effect**
An outcome that shows consistent differences between all levels of a factor.

**external validity**
The degree to which the conclusions in your study would hold for other persons in other places and at other times.

**internal validity**
The approximate truth about inferences regarding cause-effect or causal relationships.

**covariance**
A statistical estimate of the degree to which two variables vary together. This is distinct from the idea of variance which estimates the variability of a single variable.

4. *By analysis.* Statistical analysis offers you several ways to rule out alternative explanations. For instance, you could study the plausibility of an attrition or **mortality threat** by conducting a two-way factorial experimental design (see Chapter 9). One factor in this study might be the original treatment group designations (for example, program versus comparison group), while the other factor would be attrition (for example, dropout versus non-dropout group). The dependent measure could be the pretest or other available pre-program measures. A **main effect** on the attrition factor would be indicative of a threat to **external validity** or generalizability, whereas an interaction between group and attrition factors would point to a possible threat to **internal validity**. Where both effects occur, it is reasonable to infer that there is a threat to both internal and external validity.

   The plausibility of alternative explanations might also be minimized using **covariance** analysis (see the discussion of covariance in Chapter 9). For example, in a study of the effects of workfare programs on social welfare case loads, one plausible alternative explanation might be the status of local economic conditions. Here, it might be possible to construct a measure of economic conditions and include that measure as a covariate in the statistical analysis in order to adjust for or remove this factor from the outcome scores. You must be careful when using covariance adjustments of this type; perfect covariates do not exist in most social research, and the use of imperfect covariates does not completely adjust for potential alternative explanations. Nevertheless, demonstrating that treatment effects occur even after adjusting on a number of good covariates strengthens causal assertions.
5. *By preventive action.* When you anticipate potential threats, you can often rule them out by taking some type of preventive action. For example, if the program is a desirable one, it is likely that the comparison group would feel jealous or demoralized. You can take several actions to minimize the effects of these attitudes, including offering the program to the comparison group upon completion of the study or using program and comparison groups that have little opportunity for contact and communication. In addition, you can use auditing methods and quality control to track potential experimental dropouts or to insure the standardization of measurement.

These five methods for reducing the threats to internal validity should not be considered mutually exclusive. The inclusion of measurements designed to minimize threats to validity will obviously be related to the design structure and is likely to be a factor in the analysis. A good research plan should, wherever possible, make use of multiple methods for reducing threats. In general, reducing a particular threat by design or preventive action is stronger than by using one of the other three approaches. Choosing which strategy to use for any particular threat is complex and depends at least on the cost of the strategy and on the potential seriousness of the threat.

## 11-1b Building a Design

Here is where the rubber meets the road, design-wise. In the next few sections, I'll take a look at the different elements or pieces in a design and then show you how you might think about putting them together to create a tailored design to address your own research context.

### Basic Design Elements

Most research designs can be constructed from four basic elements:

1. *Time.* A causal relationship, by its very nature, implies that some time has elapsed between the occurrence of the cause and the consequent effect. Although for some phenomena, the elapsed time is measured in microseconds and is

therefore unnoticeable to a casual observer, you normally assume that the cause and effect in social science arenas do not occur simultaneously. In design notation, you indicate this temporal element horizontally. You place the symbol used to indicate the presumed cause to the left of the symbol, indicating measurement of the effect. Thus, as you read from left to right in design notation, you are reading across time. Complex designs might involve a lengthy sequence of observations and programs or treatments across time.

2. *Program(s) or treatment(s).* The presumed cause may be a program or treatment under the explicit control of the researcher or the occurrence of some natural event or program not explicitly controlled. Recall from Chapter 7 that in design notation, you usually depict a presumed cause with the symbol *X*. When multiple programs or treatments are being studied using the same design, you keep the programs distinct by using subscripts such as $X_1$ or $X_2$. For a comparison group (one that does not receive the program under study), no *X* is used.
3. *Observation(s) or measure(s).* Measurements are typically depicted in design notation with the symbol *O*. If the same measurement or observation is taken at every point in time in a design, this *O* is sufficient. Similarly, if the same set of measures is given at every point in time in this study, the *O* can be used to depict the entire set of measures. However, if you give different measures at different times, it is useful to subscript the *O* to distinguish between measurements and points in time.
4. *Groups or individuals.* The final design element consists of the intact groups or the individuals who participate in various conditions. Typically, there will be one or more program and comparison groups. In design notation, each group is indicated on a separate line. Furthermore, the manner in which groups are assigned to the conditions can be indicated by an appropriate symbol at the beginning of each line. In these cases, *R* represents a randomly assigned group, *N* depicts a nonrandomly assigned group (a nonequivalent group or cohort), and *C* indicates that the group was assigned using a cutoff score on a measurement.

Perhaps the easiest way to understand how these four basic elements become integrated into a design structure is to give several examples. One of the most commonly used designs in social research is the two-group pre-post design, which is shown in Figure 11–1.

The two lines in the design indicate that the study was composed of two groups. The two groups were nonrandomly assigned as indicated by the *N*. Both groups were measured before the program or treatment occurred as indicated by the first *O* in each line. Following this pre-observation, the group in the first line received a program or treatment, while the group in the second line did not. Finally, both groups were measured subsequent to the program.

Another common design is the posttest-only randomized experiment. The design can be depicted as shown in Figure 11–2.

Here, two groups are randomly selected, with one group receiving the program and one acting as a comparison group. Both groups are measured after the program is administered.

**Expanding a Design** With this brief review of design notation, you are now ready to understand one of the basic procedures you can use to create a tailored design—the idea of expanding basic design elements. Expanding involves combining the four basic design elements in different ways to arrive at a specific design that is appropriate for the setting at hand. As a reference or basis for all expansion, think of a design that includes only a cause and its observed effect (Figure 11–3).

This is the simplest design in causal research and serves as a starting point for the development of tailored strategies. When you add to this basic design, you are essentially expanding one of the four basic elements described previously. Each possible expansion has implications both for the cost of the study and for the

**FIGURE 11–1** Design notation for the two-group, pre-post nonequivalent-groups design

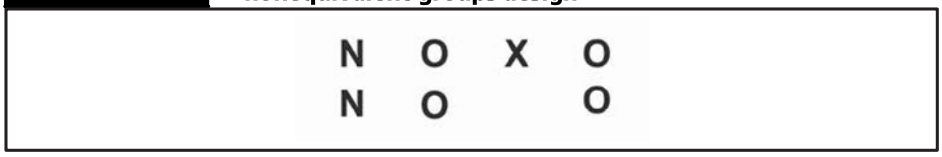



**FIGURE 11–2** The posttest-only randomized experimental design


**FIGURE 11–3** The simplest causal design with the cause and its observed effect


threats that might be ruled out. Next I will discuss the four most common ways to expand on this simple design.

**Expanding across Time** You can add to the basic design by including additional observations either before or after the program, or by adding or removing the program or different programs. For example, you might add one or more pre-program measurements and achieve the design shown in Figure 11–4.

The addition of such pretests provides a baseline that, for instance, helps assess the potential of a maturation or testing threat. If a change occurs between the first and second pre-program measures, it is reasonable to expect that similar changes might take place between the second pretest and the posttest even in the absence of the program. However, if no change occurs between the two pretests, you might more confidently assume that maturation or testing is not a likely alternative explanation for the cause-effect relationship you hypothesized. Similarly, you could add additional post-program measures, which would be useful for determining whether an immediate program effect decays over time, or whether there is a lag in time between the initiation of the program and the occurrence of an effect. You might also add and remove the program over time as shown in Figure 11–5.

The design notation in Figure 11–5 shows one form of what is sometimes called the *ABAB design* that is frequently used in clinical psychology and psychiatry. The design is particularly strong against a **history threat** because when you repeat the program, it is less likely that unique historical events are responsible for replicated outcome patterns.

**history threat**
A threat to internal validity that occurs when some historical event affects your study outcome.

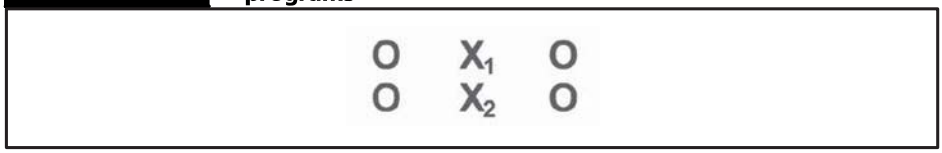
**Expanding across Programs** You have just seen that you can expand the program by adding it or removing it across time. Another way to expand the program would be to partition it into different levels of treatment. For example, in a study of the effect of a novel drug on subsequent behavior, you might use more than one dosage of the drug (see the design notation in Figure 11–6).

This design is an example of a simple factorial design with one factor having two levels. Notice that group assignment is not specified, indicating that any type of assignment might have been used. This is a common strategy in a sensitivity or parametric study, where the primary focus is on the effects obtained at various program

**FIGURE 11–4** A double-pretest single-group design created by expanding across time


**FIGURE 11–5** An add-remove design formed by expanding program and observation elements over time


**FIGURE 11–6** A two-treatment design formed by expanding across programs




levels. In a similar manner, you might expand the program by varying specific components of it across groups, which might be useful if you wanted to study different modes of the delivery of the program, different sets of program materials, and the like. Finally, you can expand the program by using theoretically polarized or opposite treatments. A comparison group is one example of such a polarization. Another might involve use of a second program that you expect to have an opposite effect on the outcome measures. A strategy of this sort provides evidence that the outcome measure is sensitive enough to differentiate between different programs.

**Expanding across Observations** At any point in time in a research design, it is usually desirable to collect multiple or redundant measurements. For example, you might add a number of measures that are similar to each other to determine whether their results converge. Or, you might want to add measurements that theoretically should not be affected by the program in question to demonstrate that the program discriminates between effects. Strategies of this type are useful for achieving convergent and discriminant validity of measures as discussed in Chapter 3. Another way to expand the observations is by proxy measurements (see Section 10-3a, The Proxy Pretest Design). Assume that you wanted to study a new educational program but neglected to take pre-program measurements. You might use a standardized achievement test for the posttest and grade point average records as a proxy measure of student achievement prior to the initiation of the program. Finally, you might also expand the observations through the use of "recollected" measures. Again, if you were conducting a study and had neglected to administer a pretest or desired information in addition to the pretest information, you might ask participants to recall how they felt or behaved prior to the study and use this information as an additional measure. Different measurement approaches obviously yield data of different quality. What is advocated here is the use of multiple measurements rather than reliance on only a single strategy.

**Expanding across Groups** Often, it will be to your advantage to add additional groups to a design to rule out specific threats to validity. For example, consider the pre-post, two-group randomized experimental design in Figure 11–7.

If this design were implemented within a single institution where members of the two groups were in contact with each other, one might expect intergroup com-

**FIGURE 11–7** The basic pre-post randomized experimental design

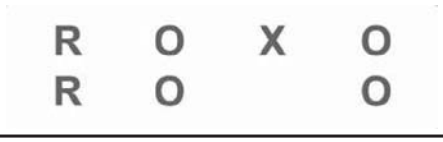



**FIGURE 11–8** A randomized experiment expanded with a nonequivalent control group


**FIGURE 11–9** A randomized experiment expanded with a nonequivalent group to help rule out a testing threat


munication, group rivalry, or demoralization of a group denied a desirable treatment or given an undesirable one to pose threats to the validity of the causal inference. (Social threats to internal validity are covered in Chapter 7.) In such a case, you might add an additional nonequivalent group from a similar institution that consists of persons unaware of the original two groups (Figure 11–8).

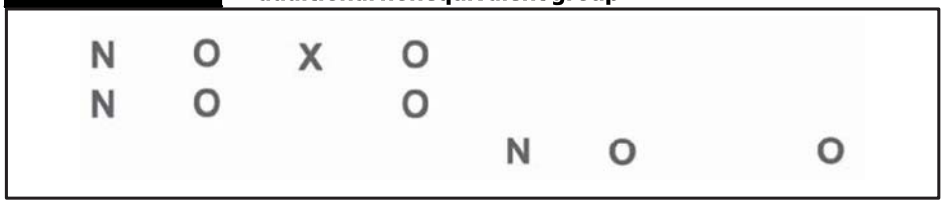
In a similar manner, whenever you use nonequivalent groups in a study it is usually advantageous to have multiple replications of each group. The use of many nonequivalent groups helps minimize the potential of a particular **selection bias** affecting the results. In some cases, it may be desirable to include the norm group as an additional group in the design. Norming group averages are available for most standardized achievement tests, for example, and might comprise an additional nonequivalent control group. You can also use cohort groups in a number of ways. For example, you might use a single measure of a cohort group to help rule out a testing threat (Figure 11–9).

**selection bias**
Any factor other than the program that leads to posttest differences between groups.

In this design, the randomized groups might be sixth graders from the same school year, and the cohort might be the entire sixth grade from the previous academic year. This cohort group did not take the pretest, and if it is similar to the randomly selected control group, it would provide evidence for or against the notion that taking the pretest had an effect on posttest scores. You might also use pre-post cohort groups (Figure 11–10).

Here, the treatment group consists of sixth graders, the first comparison group of seventh graders in the same year, and the second comparison group consists of the following year's sixth graders (the fifth graders during the study year). Strategies of this sort are particularly useful in nonequivalent designs where selection bias is a potential problem and where routinely collected institutional data is available. Finally, one other approach for expanding the groups involves partitioning groups with different assignment strategies. For example, you might randomly divide nonequivalent groups or select nonequivalent subgroups from randomly

**FIGURE 11–10** **A nonequivalent-groups design expanded with an additional nonequivalent group**




assigned groups. An example of this sort involving the combination of **random assignment** and assignment by a cutoff is discussed in detail in the following section.

**random assignment**
Process of assigning your sample into two or more subgroups by chance. Procedures for random assignment can vary from flipping a coin to using a table of random numbers to using the random number capability built into a computer.

## 11-1c A Simple Strategy for Design Construction

Considering the basic elements of a research design or the possibilities for expansion are not alone sufficient. You need to be able to integrate these elements with an overall strategy. In addition, you need to decide which potential threats are best handled by design rather than by argument, measurement, analysis, or preventive action.

While no definitive approach for designing designs exists, I suggest a tentative strategy based on the notion of expansion discussed previously. First, you begin the designing task by setting forth a design that depicts the simple hypothesized causal relationship. Second, you deliberately overexpand this basic design by expanding across time, program, observations, and groups. At this step, the emphasis is on accounting for as many likely alternative explanations or threats to validity as possible using the design. Finally, you scale back this overexpanded version considering the effect of eliminating each design component. It is at this point that you face the difficult decisions concerning the costs of each design component and the advantages of ruling out specific threats using other approaches.

Several advantages result from using this type of approach to design construction. First, you are forced to be explicit about the decisions you create. Second, the approach is conservative in nature. The strategy minimizes the chance of your overlooking a major threat to validity in constructing your design. Third, you arrive at a design that is tailored to the situation at hand. Finally, the strategy is cost-efficient. Threats you can account for by some other, less costly approach need not be accounted for in the design itself.

## 11-1d An Example of a Hybrid Design

Some of the ideas discussed in the previous sections can be illustrated in an example. To my knowledge, the design I'm presenting here has never been used, although it has strong features to commend it.

Let us assume that you want to study the effects of a new compensatory education program on subsequent student achievement. The program is designed to help students who are poor in reading improve reading skills. You can begin with the simple hypothesized cause-effect relationship (Figure 11–11).

Here, the *X* represents the reading program and the *O* stands for a reading achievement test. Assume you decide that it is desirable to add a pre-program measure so that you might investigate whether the program improves reading test scores. You also decide to expand across groups by adding a comparison group. At this point, you have the typical notation shown in Figure 11–12.

The next problem concerns how to assign the two groups. Since the program is specifically designed to help students who need special assistance in reading, you rule out random assignment because it would require denying the program to students in need. You considered the possibility of offering the program to one

**FIGURE 11–11** The cause-effect relationship: the starting point for tailoring a design


**FIGURE 11–12** A pre-post, two-group design


**FIGURE 11–13** A randomized experimental design nested within a regression-discontinuity design


randomly assigned group in the first year and to the control group in the second, but ruled that out on the grounds that it would require 2 years of program expenses and the denial of a potentially helpful program for half of the students for a period of a year. Instead, you decide to assign students by means of a cutoff score on the pretest. All students scoring below a preselected percentile on the reading pretest would be given the program while those above that percentile would act as controls. (The RD design is covered in Chapter 10.) However, experience with this strategy shows that it is difficult to adhere to a single cutoff score for assignment to a group. Of special concern is the fact that teachers or administrators might allow students who score slightly above the cutoff point into the program because they have little confidence in the ability of the achievement test to make fine distinctions in reading skills for children who score close to the cutoff. To deal with this potential problem, you decide to partition the groups using a particular combination of assignment by a cutoff and random assignment as shown in Figure 11–13.

This design has two cutoff points. All those scoring below a certain percentile are assigned to the program group automatically by this cutoff. All those scoring above another higher percentile are automatically assigned to the comparison group by this cutoff. Finally, all those who fall in the interval between the cutoffs on the pretest are randomly assigned to either the program or comparison group.

This strategy has several advantages. It directly addresses the concern to teachers and administrators that the test may not be able to discriminate well between students who score immediately above or below a cutoff point. For example, a student whose true ability in reading would place him or her near the cutoff might have a bad day and therefore might be placed into the treatment or comparison group by chance factors. The design outlined in Figure 11–13 is defensible. You can agree with the teachers and administrators that the test is fallible. Nevertheless, since you need some criteria to assign students to the program, you can argue that the fairest approach would be to assign borderline cases by lottery. In addition, by combining two excellent strategies (the randomized experiment and the regression-

discontinuity) you can analyze results separately for each and address the possibility that design factors might bias results.

Many other worthwhile strategies are not mentioned in the previous scenario. For example, instead of using simple randomized assignment within the cutoff interval, you might use a weighted random assignment so that students scoring lower in the interval have a greater probability of being assigned to the program. In addition, you might consider expanding the design in a number of other ways, by including double pretests or multiple posttests; multiple measures of reading skills; additional replications of the program or variations of the programs and additional groups, such as norming groups; controls from other schools; and the like. Nevertheless, this brief example serves to illustrate the advantages of explicitly constructing a research design to meet the specific needs of a particular situation.

## 11-1e The Nature of Good Design

Throughout the design construction task, it is important to have in mind some endpoint—some criteria that you should try to achieve before finally accepting a design strategy. The criteria discussed in the following sections are only meant to be suggestive of the characteristics found in good research design. It is worth noting that all of these criteria point to the need to individually tailor research designs rather than accepting standard textbook strategies as is.

- *Theory-grounded.* Good research strategies reflect the theories that you are investigating. When you hypothesize specific theoretical expectations, you should then incorporate them into the design. For example, when theory predicts a specific treatment effect on one measure but not on another, the inclusion of both in the design improves discriminant validity and demonstrates the predictive power of the theory.
- *Situational.* Good research designs reflect the settings of the investigation. This was illustrated in the previous section where a particular need of teachers and administrators was explicitly addressed in the design strategy. Similarly, you can assess intergroup rivalry, demoralization, and competition through the use of additional comparison groups not in direct contact with the original group.
- *Feasible.* Good designs can be implemented. You must carefully plan the sequence and timing of events. You need to anticipate potential problems in measurement, adherence to assignment, database construction, and the like. Where needed, you should include additional groups or measurements in the design to explicitly correct for such problems.
- *Redundant.* Good research designs have some flexibility built into them. Often, this flexibility results from duplication of essential design features. For example, multiple replications of a treatment help ensure that failure to implement the treatment in one setting will not invalidate the entire study.
- *Efficient.* Good designs strike a balance between redundancy and the tendency to overdesign. Where it is reasonable, other, less costly strategies for ruling out potential threats to validity are used.

This is by no means an exhaustive list of the criteria by which to judge good research design. Nevertheless, goals of this sort help guide you toward a final design choice and emphasize important components that should be included.

The development of a theory of research methodology for the social sciences has largely occurred over the past half century and most intensively within the past two decades. It is not surprising that in such a relatively recent effort, an emphasis on a few standard research designs has occurred. Nevertheless, by moving away from the notion of design selection and toward an emphasis on design construction, there is much to be gained in our understanding of design principles and in the quality of our research.

## 11-2 Relationships among Pre-Post Designs

Now that you are getting more sophisticated in understanding the idea of research design, you are ready to think more methodologically about some of the underlying principles that cut across design types. Here I show how the most frequently used design structures can be understood in relation to one another (Figure 11–14).

There are three major types of pre-post program-comparison group designs all sharing the basic design structure shown in Figure 11–14:

- The **RE design**
- The **NEGD design**
- The **RD design**

**RE design**
The Randomized Experimental (RE) Design is characterized by one essential feature: the random assignment of participants to conditions.

**NEGD design**
A pre-post twogroup quasi-experimental design structured like a pretest-posttest randomized experiment, but lacking random assignment to group.

**RD design**
A pretest-posttest, program-comparison group quasi-experimental design in which a cutoff criterion on the preprogram measure is the method of assignment to group.

The designs differ only in the method by which participants are assigned to the two groups. In the RE, participants are assigned randomly. In the RD design, they are assigned using a cutoff score on the pretest. In the NEGD, assignment of participants is not explicitly controlled; they might self-select into either group, or other unknown or unspecified factors might determine assignment.

Because these three designs differ so critically in their assignment strategy, they are often considered distinct or unrelated. However, it is useful to look at them as forming a continuum, both in terms of assignment and in terms of their strength with respect to internal validity.

You can look at the similarity of the three designs in terms of their assignment by graphing their assignment functions with respect to the pretest variable. In Figure 11–15, the vertical axis is the probability that a specific unit (such as a person) will be assigned to the treatment group. These values, because they are probabilities, range from 0 to 1. The horizontal axis is an idealized pretest score. Each line on the graph is an assignment function for a design.

Let's first examine the assignment function for the simple pre-post randomized experiment. Because units are assigned randomly, the probability that a unit will be assigned to the treatment group is always 1/2 or .5 (assuming equal assignment probabilities are used). This function is indicated by the horizontal red line at .5 in the figure. For the RD design, I've arbitrarily set the cutoff value at the midpoint of the pretest variable and assigned units scoring below that value to the treatment and those scoring at or above that value to the control condition. (The arguments made here would generalize to the case of high-scoring treatment cases as well.) In this case, the assignment function is a simple step function, with the probability of assignment to the treatment = 1 for the pretest scores below the cutoff and = 0 for those above. It is important to note that for both the RE and RD designs, it is an easy matter to plot assignment functions because assignment is explicitly controlled. This is not the case for the NEGD. Here, the idealized assignment function differs depending on the degree to which the groups are nonequivalent on the pretest. If they are extremely nonequivalent (with the treatment group scoring lower on the pretest), the assignment function would approach the step function of the RD design. If the groups are hardly nonequivalent at all, the function would approach the flat-line function of the randomized experiment.

The graph of assignment functions points out an important issue about the relationships among these designs: The designs are not distinct with respect to their assignment functions; they form a continuum (Figure 11–16). On one end of the continuum is the RE design and at the other is the RD. The NEGD can be viewed as a degraded RD or RE depending on whether the assignment function more closely approximates one or the other.

You can also view the designs as differing with respect to the degree to which they generate a pretest difference between the groups.

**FIGURE 11–14** **The basic pre-post, two-group design structure**

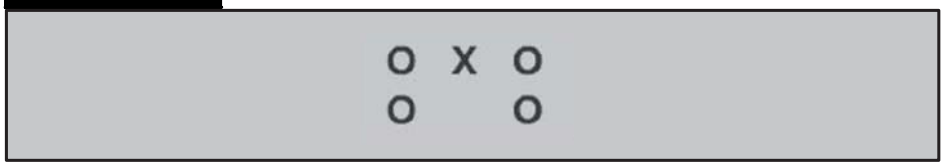



**FIGURE 11–15** **Probability of assignment to treatment for the RE, NEGD, and RD design**

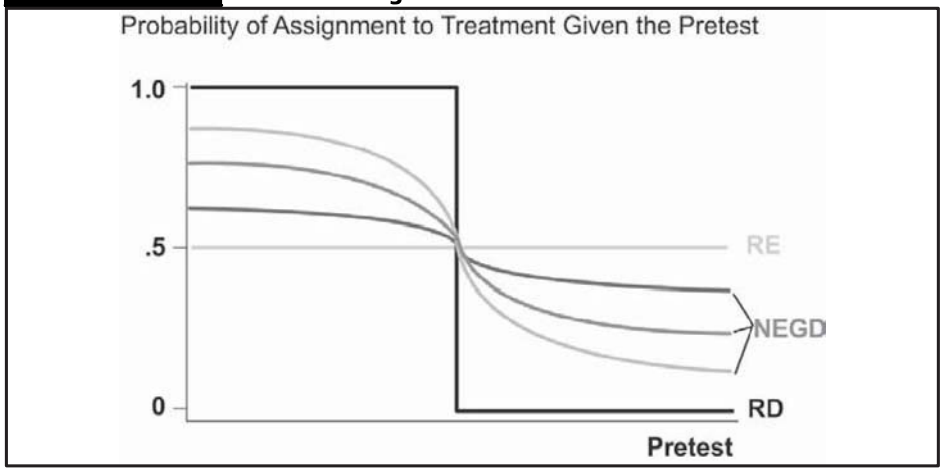



**FIGURE 11–16** **The continuum of pre-post, two-group designs in terms of preprogram equivalence**


Figure 11–16 shows that the RD design induces the maximum possible pretest difference. The RE design induces the smallest pretest difference (the most equivalent). The NEGD fills in the gap between these two extreme cases. If the groups are extremely nonequivalent, the design is closer to the RD design. If they're extremely similar, it's closer to the RE design.

Finally, you can also distinguish the three designs in terms of the *a priori* knowledge they give about assignment. It should be clear that in the RE design you know perfectly the probability of assignment to treatment; it is .5 for each participant. Similarly, with the RD design, you also know perfectly the probability of assignment. In this case, it is precisely dependent on the cutoff assignment rule. It is dependent on the pretest where the RE design is not. In both these designs, you know the assignment function perfectly, and it is this knowledge that enables you to obtain unbiased estimates of the treatment effect with these designs. This is why I conclude that, with respect to internal validity, the RD design is as strong as the RE design. With the NEGD, however, you do not know the assignment function perfectly. Because of this, you need to model this function either directly or indirectly (for example, through reliability corrections).

The major point is that you should not look at these three designs as entirely distinct. They are related by the nature of their assignment functions and the degree of pretest nonequivalence between groups. This continuum has important implications for understanding the statistical analyses of these designs. (To learn more about statistical analysis, read Chapter 12.)

# 11-3 Contemporary Issues in Research Design[2]

It is fitting to end this section on research design by reflecting on where the research design endeavor stands today and trying to identify what the major, cutting-edge issues in design currently are. Of course, cutting-edge issues can always turn into a two-edged sword, so let's be careful!

Research design is a relatively recent invention. It really didn't exist in any formal sense prior to the 20th century and didn't really become delineated until the 1950s and 1960s. In the last half of the 20th century, this area has primarily involved explication of two interrelated topics: the theory of the validity of casual inferences and a taxonomy of the research designs that allow the examination of causal hypotheses.

Here I want to make the case that in the past decade traditional thinking has moved beyond the traditional thinking about design as simply a collection of specific designs and threats to validity has been replaced with a more integrated, synthetic view of design as part of a general logical and epistemological framework for research. To support this view that the notion of research design is evolving toward increasing integration, I will present a number of themes that seem to characterize current thinking and that cut across validity typologies and design taxonomies. This list of themes may also be viewed as a tentative description of the advances in thinking about research design in social research.

## 11-3a The Role of Judgment

One theme that underlies most of the others and that illustrates the increasing awareness of the tentativeness and frailty of research concerns the importance of human judgment in research. I know it's probably obvious to you that the personal subjective judgments of researchers has a major effect on research, but believe it or not, I and a lot of my colleagues really seemed to lose sight of this fact over the past half century and are only rediscovering it now. We were lured by the idea that we might be able to mechanize the research design process, to automate it in a sense to minimize the role of human judgment and (we thought) improve the objectivity of our work. But I think we now realize that objectivity (at least in the old-fashioned positivist sense) isn't all it was cracked up to be and that it is heavily dependent on human judgment itself. Researchers are beginning to think about the psychological components of cause-effect relationships and causal reasoning and are increasingly incorporating models of the judgmental process into their research designs and analyses. And researchers are also recognizing more clearly the sociological bases of scientific thought and the fact that science is at root a human enterprise. We increasingly recognize that scientific communities have social norms and customs and often operate like tribal groups (and, unfortunately, are sometimes as primitive). The positivist, mechanistic view is all but gone from contemporary design thinking, and what remains is a more judgmental and ironically, a more scientifically sensible perspective.

[2]Parts of this section were based on Trochim, W. (Ed.), (1986). Editor's Notes. *Advances in Quasi-Experimental Design and Analysis. New Directions for Program Evaluation Series*, Number 31, San Francisco, CA: Jossey-Bass.

## 11-3b The Case for Tailored Designs

In the early days, methodologists took a taxonomic approach to design, laying out a collection of relatively discrete research designs and discussing how weak or strong they were for valid causal inference. Presentations of research designs were full of discussions of classification issues and specialized design notation systems (a lot like the *X* and *O* system of notation I present here and which my students fondly refer to as the tic-tac-toe school of design notation). Almost certainly, these early design proponents recognized that there was a virtual infinity of design variations and that validity was more complexly related to theory and context than their presentations implied. Nonetheless, what seemed to evolve was a cookbook approach to design that largely involved picking a design off the shelf and checking off lists of validity threats.

In the past few decades, we've gotten a little more sophisticated than that, constructing tailored research designs as combinations of more elemental units (for example, assignment strategies, measurement occasions) based on the specific contextual needs and plausible alternative explanations for a treatment effect (as described in the first section in this chapter). The implication for you is that you should focus on the advantages of different combinations of design features rather than on a relatively restricted set of prefabricated designs. In writing this text, I try (without always succeeding) to encourage you to break away from this canned, off-the-shelf, taxonomic design mentality, and I emphasize design principles and issues that cut across the traditional distinctions between true experiments, nonexperiments, and quasi-experiments (as in the discussion of the previous section of this chapter).

## 11-3c The Crucial Role of Theory

Research design has sometimes been criticized for encouraging an atheoretical, black-box research mentality. People are assigned to either complex, convoluted programs, or (often) to equally complex comparison conditions. The machinery of random assignment (or our quasi-experimental attempts to approximate random assignment) are the primary means of defining whether the program has an effect. If you think about it, this comparison group mentality is inherently atheoretical and noncontextual. It assumes that the same design mechanism works in exactly the same way whether you apply it in studies of mental health, criminal justice, income maintenance, or education.

There is nothing inherently wrong with this program-group-versus-comparison-group logic. The problem is that it may be a rather crude, uninformative approach. In the two-group case, you are simply creating a dichotomous input into reality. If you observe a posttest difference between groups, it could be explained by this dichotomous program-versus-comparison-group input or by any number of alternative explanations, including differential attrition rates, intergroup rivalry and communication, initial selection differences among groups, or different group histories. Researchers usually try to deal with these alternative explanations by ruling them out through argument, additional measurement, patched-up design features, and auxiliary analysis.

But we now see that there may be another way to approach research that emphasizes theoretical explanation more and simplistic design structure less. For instance, we have begun to emphasize greater use of patterns in research by using more complex theory-driven predictions that, if corroborated, allow fewer plausible alternative explanations for the effect of a program. (**Pattern matching** is covered in Chapter 10.) Because appropriate theories may not be readily available, especially for the evaluation of contemporary social programs, we are developing methods and processes to help people articulate the implicit theories that program administrators and stakeholder groups have in mind and which presumably guide the formation and implementation of the program.

**pattern matching**
The degree of correspondence between two data items. For instance, you might look at a pattern match of a theoretical expectation pattern with an observed patten to see if you are getting the outcomes you expect.

## 11-3d Attention to Program Implementation

A theory-driven approach to research will be futile unless we can demonstrate that the program was in fact carried out or implemented as the theory intended. I know this is obvious to you, but once again, it's astonishing how often people like me forget these basic truths. In the past few decades, we have seen the development of program implementation theory that looks at the process of program execution as an important part of research itself. For instance, one approach emphasizes the development of organizational procedures and training systems that accurately transmit the program and that anticipate likely institutional sources of resistance. Another strategy involves the assessment of program delivery through program audits, management information systems, and the like. This emphasis on program implementation has further obscured the traditional distinction between process and outcome evaluation. At the least, it is certainly clear that good research cannot be accomplished without attending to program processes, and we are continuing to develop better notions of how to combine these two efforts.

## 11-3e The Importance of Quality Control

Over and over, our experience with research has shown that even the best-laid research plans often go awry in practice, sometimes with disastrous results. Okay, I know this is another one of those things that should have been obvious, but at least we're finally beginning to catch on now. Over the past decade, researchers have begun to pay increasing attention to the integrity and quality of research designs in real-world settings. One way to do this is to go to people who know something about data integrity and quality assurance and incorporate techniques used by these other professions: accounting, auditing, industrial quality control. For instance, double-bookkeeping can be used to keep verifiable records of research participation. Acceptance sampling can be an efficient method for checking accuracy in large data collection efforts where an exhaustive examination of records is impractical or excessive in cost. These issues are particularly important in quasi-experimental research design, where it is especially important to demonstrate that sampling, measurement, group assignment, and analysis decisions do not interact with program participation in ways that can confound the final interpretation of results.

## 11-3f The Advantages of Multiple Perspectives

Researchers have long recognized the importance of replication and systematic variation in research. In the past few years, we have rediscovered this principle. (There does seem to be an awful lot of rediscovering going on in this discussion, doesn't there?) The emphasis on multiple perspectives rests on the notion that no single point of view will ever be sufficient for understanding a phenomenon with validity. Multiple realizations—of research questions, measures, samples, designs, analyses, replications, and so on—are essential for convergence on the truth (and even then we're lucky if we get there). However, such a varied approach can become a methodological and epistemological Pandora's Box unless researchers apply critical judgment in deciding which multiples to use in a study or set of studies. That's the challenge, and researchers are only beginning to address it.

## 11-3g Evolution of the Concept of Validity

The history of research design is inseparable from the development of the theory of the validity of causal inference. For decades researchers have been arguing about the definition of validity and debating whether it's more important to the establishment of a cause-effect relationship (internal validity) or whether we should

emphasize generalizability (external validity). Some researchers argued that it was more important to nail down the cause-effect relationship even for nonrepresentative people in one place at one time and then worry about generalizing in subsequent studies that attempt to replicate the original study. Others worried that it doesn't make sense to pour our resources into intense rigorous studies of a particular group in one place and at one time because this has no generalizability and little policy relevance. Believe it or not, I remember having numerous intense debates about this dilemma as a graduate student. Of course, the obvious solution—that we want to achieve a balance between internal and external validity, between establishing the cause-effect relationship with precision and sampling broadly enough to have some generalizability—has emerged with painful slowness over time. But, at least we got there. More and more, research design is seen as a balancing act, using judgment to allocate precious and scarce resources to blend different levels of validity.

### 11-3h Development of Increasingly Complex Realistic Analytic Models

In the past decade, researchers have made considerable progress toward complicating statistical analyses to account for increasingly complex contexts and designs. For all of you who have to take statistics this is, of course, the bad news. In the past 50 years, we have developed more complex statistics for dealing with measurement error, creating dichotomous dependent variables, estimating invisible traits and characteristics, and so on. In fact, I think that many of these advances are among the most important contributions social science has made in the past 50 years. Too bad so few people understand them! Of course, it would help if we could learn how to teach statistics—especially the newer, more complex, and exciting approaches—to real people. Who knows, maybe we'll accomplish that in the next century. In the meantime, those of you who have to slog through advanced stats courses can perhaps take heart that the complexity you're grappling with actually represents a legitimate advance.

Parallel to the development of these increasingly complex, realistic statistical models, cynicism has deepened among researchers about the ability of any single model or analysis to be sufficient. (And that's really saying something because researchers started out as a cynical crowd.) Increasingly researchers are calling for multiple statistical analyses and using the results to bracket the likely true estimates. Researchers have virtually abandoned hope of finding a single correct analysis and have accordingly moved to multiple analyses that are based on systematically distinct assumptional frameworks and that rely in an increasingly direct way on the role of judgment.

## Summary

So, where does this leave all of us who do social research? The good news is that all of these advances suggest that researchers have become much more realistic about what research can accomplish. Gone are the heady days of the 1960s and 1970s where we hoped to be able to turn applied social research into a branch of science akin to physics or chemistry. The bad news is that this makes our lives considerably more complicated. Researchers have discovered a lot of problems in our initial approaches to social research and we've invented ever more complicated solutions for them. The overall picture that emerges about contemporary research is that research design is judgmental. It is based on multiple and varied sources of evidence; it should be multiplistic in realization; it must attend to process as well as to outcome; it is better off when theory driven; and it leads ultimately to multiple analyses that attempt to bracket the program effect within some reasonable range.

In one sense, this is hardly a pretty picture. Contemporary views about research design and its role in causal inference are certainly more tentative and critical than they were in

1965 or perhaps even in 1979. But, this more integrated and complex view of research has emerged directly from our experiences in the conduct of such studies. Perhaps the social research community is learning how to do this stuff better. At least, that's the hope.

Login to the Online Edition of your text at www.atomicdog.com to find additional resources located in the Study Guide at the end of each chapter.