2

Experimental Design

Roger E. Kirk

SOME BASIC DESIGN CONCEPTS

Sir Ronald Fisher, the statistician, eugenicist, evolutionary biologist, geneticist, and father of modern experimental design, observed that experiments are ‘only experience carefully planned in advance, and designed to form a secure basis of new knowledge’ (Fisher, 1935: 8). Experiments are characterized by the: (1) manipulation of one or more independent variables; (2) use of controls such as randomly assigning participants or experimental units to one or more independent variables; and (3) careful observation or measurement of one or more dependent variables. The first and second characteristics—manipulation of an independent variable and the use of controls such as randomization—distinguish experiments from other research strategies.

The emphasis on experimentation in the sixteenth and seventeenth centuries as a way of establishing causal relationships marked the emergence of modern science from its roots in natural philosophy (Hacking, 1983). According to nineteenth-century philosophers, a causal relationship exists: (1) if the cause precedes the effect; (2) whenever the cause is present, the effect occurs; and (3) the cause must be present for the effect to occur. Carefully designed and executed experiments continue to be one of science’s most powerful methods for establishing causal relationships.

Experimental design

An *experimental design* is a plan for assigning experimental units to treatment levels and the statistical analysis associated with the plan (Kirk, 1995: 1). The design of an experiment involves a number of inter-related activities.

1. Formulation of statistical hypotheses that are germane to the scientific hypothesis. A statistical hypothesis is a statement about: (a) one or more parameters of a population or (b) the functional form of a population. Statistical hypotheses are rarely identical to scientific hypotheses—they are testable formulations of scientific hypotheses.
2. Determination of the treatment levels (independent variable) to be manipulated, the measurement to be recorded (dependent variable), and the extraneous conditions (nuisance variables) that must be controlled.
3. Specification of the number of experimental units required and the population from which they will be sampled.
4. Specification of the randomization procedure for assigning the experimental units to the treatment levels.

5. Determination of the statistical analysis that will be performed (Kirk, 1995: 1–2).

In summary, an experimental design identifies the independent, dependent, and nuisance variables and indicates the way in which the randomization and statistical aspects of an experiment are to be carried out. The primary goal of an experimental design is to establish a causal connection between the independent and dependent variables. A secondary goal is to extract the maximum amount of information with the minimum expenditure of resources.

Randomization

The seminal ideas for experimental design can be traced to Sir Ronald Fisher. The publication of Fisher's *Statistical methods for research workers* in 1925 and *The design of experiments* in 1935 gradually led to the acceptance of what today is considered the cornerstone of good experimental design: randomization. Prior to Fisher's pioneering work, most researchers used systematic schemes rather than randomization to assign participants to the levels of a treatment. Random assignment has three purposes. It helps to distribute the idiosyncratic characteristics of participants over the treatment levels so that they do not selectively bias the outcome of the experiment. Also, random assignment permits the computation of an unbiased estimate of error effects—those effects not attributable to the manipulation of the independent variable—and it helps to ensure that the error effects are statistically independent. Through random assignment, a researcher creates two or more groups of participants that at the time of assignment are probabilistically similar on the average.

Quasi-experimental design

Sometimes, for practical or ethical reasons, participants cannot be randomly assigned to treatment levels. For example, it would be unethical to expose people to a disease to evaluate the efficacy of a treatment. In such cases it may be possible to find preexisting or naturally occurring experimental units who have been exposed to the disease. If the research has all of the features of an experiment except random assignment, it is called a *quasi-experiment*. Unfortunately, the interpretation of quasi-experiments is often ambiguous. In the absence of random assignment, it is difficult to rule out all variables other than the independent variable as explanations for an observed result. In general, the difficulty of unambiguously interpreting the outcome of research varies inversely with the degree of control that a researcher is able to exercise over randomization.

Replication and local control

Fisher popularized two other principles of good experimentation: replication and local control or blocking. Replication is the observation of two or more experimental units under the same conditions. Replication enables a researcher to estimate error effects and obtain a more precise estimate of treatment effects. Blocking, on the other hand, is an experimental procedure for isolating variation attributable to a nuisance variable. Nuisance variables are undesired sources of variation that can affect the dependent variable. Three experimental approaches are used to deal with nuisance variables:

1. Hold the variable constant.
2. Assign experimental units randomly to the treatment levels so that known and unsuspected sources of variation among the units are distributed over the entire experiment and do not affect just one or a limited number of treatment levels.
3. Include the nuisance variable as one of the factors in the experiment.

The latter approach is called *local control* or *blocking*. Many statistical tests can be thought of as a ratio of error effects and treatment effects as follows:

$$\text{Test statistic} = \frac{f(\text{error effects}) + f(\text{treatment effects})}{f(\text{error effects})} \qquad (1)$$

where $f()$ denotes a function of the effects in parentheses. Local control, or blocking, isolates variation attributable to the nuisance variable so that it does not appear in estimates of error effects. By removing a nuisance variable from the numerator and denominator of the test statistic, a researcher is rewarded with a more powerful test of a false null hypothesis.

Analysis of covariance

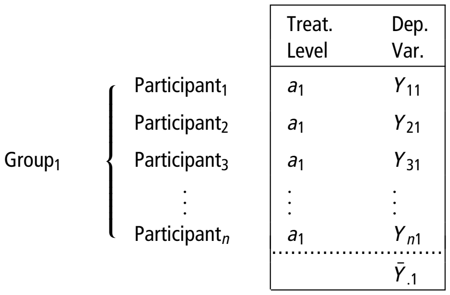
A second general approach can be used to control nuisance variables. The approach, which is called *analysis of covariance*, combines regression analysis with analysis of variance. Analysis of covariance involves measuring one or more concomitant variables in addition to the dependent variable. The concomitant variable represents a source of variation that has not been controlled in the experiment and one that is believed to affect the dependent variable. Through analysis of covariance, the dependent variable is statistically adjusted to remove the effects of the uncontrolled source of variation.

The three principles that Fisher vigorously championed—randomization, replication, and local control—remain the foundation of good experimental design. Next, I describe some threats to internal validity in simple experimental designs.

THREATS TO INTERNAL VALIDITY IN SIMPLE EXPERIMENTAL DESIGNS

One-group posttest-only design

One of the simplest experimental designs is the one-group posttest-only design (Shadish, et al., 2002: 106–107). A diagram of the design is shown in Figure 2.1. In the design, a sample of participants is exposed to a treatment after which the dependent variable is measured. I use the terms ‘treatment,’ ‘factor,’ and ‘independent variable’ interchangeably. The treatment has one level denoted by a_1. The design does not have a control group or a pretest. Hence, it is necessary to compare the dependent-variable mean, $\bar{Y}_{.1}$, with what the researcher thinks would happen in the absence of the treatment.



		Treat. Level	Dep. Var.
Group_1	Participant_1	a_1	Y_{11}
	Participant_2	a_1	Y_{21}
	Participant_3	a_1	Y_{31}
	⋮	⋮	⋮
	Participant_n	a_1	Y_{n1}
			$\bar{Y}_{.1}$

Figure 2.1 Layout for a one-group posttest-only design. The $i = 1, \ldots, n$ participants receive the treatment level (Treat. Level) denoted by a_1 after which the dependent variable (Dep. Var.) denoted by Y_{i1} is measured. The mean of the dependent variable is denoted by $\bar{Y}_{.1}$.

It is difficult to draw unambiguous conclusions from the design because of serious threats to internal validity. *Internal validity* is concerned with correctly concluding that an independent variable is, in fact, responsible for variation in the dependent variable. One threat to internal validity is *history*: events other than the treatment that occur between the time the treatment is presented and the time that the dependent variable is measured. Such events, called *rival hypotheses*, become more plausible the longer the interval between the treatment and the measurement of the dependent variable. Another threat is *maturation*. The dependent variable may reflect processes unrelated to the treatment that occur simply as a function of the passage of time: growing older, stronger, larger, more experienced, and so on. *Selection* is another serious threat to the internal validity of the design. It is possible that the participants in the experiment are different from those in the hypothesized comparison sample. The one-group posttest-only design should only be used in situations where the researcher is able to accurately specify the value of the mean that would be observed in the absence of the treatment. Such situations are rare in the behavioral sciences and education.

		Dep. Var.	Treat. Level	Dep. Var.
Group$_1$	Block$_1$	Y_{11}	a_1	Y_{12}
	Block$_2$	Y_{21}	a_1	Y_{22}
	Block$_3$	Y_{31}	a_1	Y_{32}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	Block$_n$	Y_{n1}	a_1	Y_{n1}
		$\bar{Y}_{.1}$		$\bar{Y}_{.2}$

Figure 2.2 Layout for a one-group pretest-posttest design. The dependent variable (Dep. Var.) for each of the *n* blocks of participants is measured prior to the presentation of the treatment level, a_1, and again after the treatment level (Treat. Level) has been presented.The means of the pretest and posttest are denoted by $\bar{Y}_{.1}$ and $\bar{Y}_{.2}$, respectively.

One-group pretest-posttest design

The one-group pretest-posttest design shown in Figure 2.2 also has one treatment level, a_1. However, the dependent variable is measured before and after the treatment level is presented. The design enables a researcher to compute a contrast between means in which the pretest and posttest means are measured with the same precision. Each block in the design can contain one participant who is observed two times or two participants who are matched on a relevant variable. Alternatively, each block can contain identical twins or participants with similar genetic characteristics. The essential requirement is that the variable on which participants are matched is correlated with the dependent variable. The null and alternative hypotheses are

$$H_0 : \mu_1 - \mu_2 = \delta_0 \qquad (2)$$

$$H_1 : \mu_1 - \mu_2 \neq \delta_0, \qquad (3)$$

where δ_0 is usually equal to 0. A *t* statistic for dependent samples typically is used to test the null hypothesis.

The design is subject to several of the same threats to internal validity as the one-group posttest-only design, namely, history and maturation. These two threats become more plausible explanations for an observed difference between $\bar{Y}_{.1}$ and $\bar{Y}_{.2}$ the longer the time interval between the pretest and posttest. Selection, which is a problem with the one-group posttest-only design, is not a problem. However, as I describe next, the use of a pretest introduces new threats to internal validity: *testing*, *statistical regression*, and *instrumentation*.

Testing is a threat to internal validity because the pretest can result in familiarity with the testing situation, acquisition of information that can affect the dependent variable, or sensitization to information or issues that can affect how the dependent variable is perceived. A pretest, for example, may sensitize participants to a topic, and, as a result of focusing attention on the topic, enhance the effectiveness of a treatment. The opposite effect also can occur. A pretest may diminish participants' sensitivity to a topic and thereby reduce the effectiveness of a treatment.

Statistical regression is a threat when the mean-pretest scores are unusually high or low and measurement of the dependent variable is not perfectly reliable. Statistical regression operates to increase the scores of participants on the posttest if the mean-pretest score is unusually low and decrease the scores of participants if the mean pretest score is unusually high. The amount of statistical regression is inversely related to the reliability of the measuring instrument.

Another threat is *instrumentation*: changes in the calibration of measuring instruments between the pretest and posttest, shifts in the criteria used by observers or scorers, or unequal intervals in different ranges of measuring instruments.

The one-group posttest-only design is most useful in laboratory settings where the time interval between the pretest and posttest is short. The internal validity of the design can be improved by the addition of a second pretest as I show next.

One-group double-pretest posttest design

The plausibility of maturation and statistical regression as threats to internal validity of

		Dep. Var.	Dep. Var.	Treat. Level	Dep. Var.
Group_1	Block_1	Y_{11}	Y_{12}	a_1	Y_{13}
	Block_2	Y_{21}	Y_{22}	a_1	Y_{23}
	Block_3	Y_{31}	Y_{32}	a_1	Y_{33}
	⋮	⋮	⋮	⋮	⋮
	Block_n	Y_{n1}	Y_{n2}	a_1	Y_{n3}
		$\bar{Y}_{.1}$	$\bar{Y}_{.2}$		$\bar{Y}_{.3}$

Figure 2.3 Layout for a one-group double-pretest-posttest design. The dependent variable (Dep. Var.) for each of the n blocks of participants is measured twice prior to the presentation of the treatment level (Treat. Level), a_1, and again after the treatment level has been presented. The means of the pretests are denoted by $\bar{Y}_{.1}$ and $\bar{Y}_{.2}$; the mean of the posttest is denoted by $\bar{Y}_{.3}$.

an experiment can be reduced by having two pretests as shown in Figure 2.3. The time intervals between the three measurements of the dependent variable should be the same. The null and alternative hypotheses are, respectively,

$$H_0 : \mu_2 = \mu_3 \qquad (4)$$

$$H_1 : \mu_2 \neq \mu_3 \qquad (5)$$

The data are analyzed by means of a randomized block analysis of covariance design where the difference score, $D_i = Y_{i1} - Y_{i2}$, for the two pretests is used as the covariate. The analysis statistically adjusts the contrast $\bar{Y}_{.2} - \bar{Y}_{.3}$ for the threats of maturation and statistical regression. Unfortunately, the use of two pretests increases the threat of another rival hypothesis: testing. History and instrumentation also are threats to internal validity.

SIMPLE-EXPERIMENTAL DESIGNS WITH ONE OR MORE CONTROL GROUPS

Independent samples t-statistic design

The inclusion of one or more control groups in an experiment greatly increases the internal validity of a design. One such design is the randomization and analysis plan that is used with a t statistic for independent samples. The design is appropriate for experiments in which N participants are randomly assigned to treatment levels a_1 and a_2 with n_1 and n_2 participants in the respective levels. A diagram of the design is shown in Figure 2.4.

		Treat. Level	Dep. Var.
Group_1	Participant_1	a_1	Y_{11}
	Participant_2	a_1	Y_{21}
	⋮	⋮	⋮
	Participant_{n_1}	a_1	$Y_{n_1 1}$
			$\bar{Y}_{.1}$
Group_2	Participant_1	a_2	Y_{12}
	Participant_2	a_2	Y_{22}
	⋮	⋮	⋮
	Participant_{n_2}	a_2	$Y_{n_2 2}$
			$\bar{Y}_{.2}$

Dep. Var., dependent variable; Treat. Level, treatment level.

Figure 2.4 Layout for an independent samples t-statistic design. Twenty participants are randomly assigned to treatment levels a_1 and a_2 with $n_1 = n_2 = 10$ in the respective levels. The means of the treatment levels are denoted by $\bar{Y}_{.1}$ and $\bar{Y}_{.2}$, respectively.

Consider an experiment to evaluate the effectiveness of a medication for helping smokers break the habit. The treatment levels are a_1 = a medication delivered by a patch that is applied to a smoker's back and a_2 = a placebo, a patch without the medication. The dependent variable is a participant's rating on a ten-point scale of his or her desire for a cigarette after six months of treatment. The null and alternative hypotheses for the experiment are, respectively,

$$H_0 : \mu_1 - \mu_2 = \delta_0 \qquad (6)$$

$$H_1 : \mu_1 - \mu_2 \neq \delta_0 \qquad (7)$$

where μ_1 and μ_2 denote the means of the respective populations and δ_0 usually is equal to 0. A t statistic for independent samples is used to test the null hypothesis.

Assume that $N = 20$ smokers are available to participate in the experiment. The researcher assigns $n = 10$ smokers to each treatment level so that each of the $(np)!/(n!)^p = 184,756$ possible assignments has the same probability. This is accomplished by numbering the smokers from 1 to 20 and drawing numbers from a random numbers table. The smokers corresponding to the first 10 unique numbers drawn between 1 and 20 are assigned to treatment level a_1; the remaining 10 smokers are assigned to treatment level a_2.

Random assignment is essential for the internal validity of the experiment. Random assignment helps to distribute the idiosyncratic characteristics of the participants over the two treatment levels so that the characteristics do not selectively bias the outcome of the experiment. If, for example, a disproportionately large number of very heavy smokers was assigned to either treatment level, the evaluation of the treatment could be compromised. Also, it is essential to measure the dependent variable at the same time and under the same conditions. If, for example, the dependent variable is measured in different testing room, irrelevant events in one of the rooms—distracting outside noises, poor room air circulation, and so on—become rival hypotheses for the difference between $\bar{Y}_{.1}$ and $\bar{Y}_{.2}$.

Dependent samples t-statistic design

Let us reconsider the cigarette smoking experiment. It is reasonable to assume that the difficulty in breaking the smoking habit is related to the number of cigarettes that a person smokes per day. The design of the smoking experiment can be improved by isolating this nuisance variable. The nuisance variable can be isolated by using the experimental design for a dependent samples t statistic. Instead of randomly assigning 20 participants to the two treatment levels, a researcher can form pairs of participants who are matched with respect to the number of cigarettes smoked per day. A simple way to match the participants is to rank them in terms of the number of cigarettes they smoke. The participants ranked 1 and 2 are assigned to block one, those ranked 3 and 4 are assigned to block two, and so on. In this example, 10 blocks of matched participants can be formed. The participants in each block are then randomly assigned to treatment level a_1 or a_2. The layout for the experiment is shown in Figure 2.5. The null and alternative hypotheses for the experiment are, respectively:

	Dep. Var.	Dep. Var.	Treat. Level	Dep. Var.
Block$_1$	a_1	Y_{11}	a_2	Y_{12}
Block$_2$	a_1	Y_{21}	a_2	Y_{22}
Block$_3$	a_1	Y_{31}	a_2	Y_{32}
⋮	⋮	⋮	⋮	⋮
Block$_{10}$	a_1	$Y_{10,1}$	a_2	$Y_{10,2}$
		$\bar{Y}_{.1}$		$\bar{Y}_{.2}$

Dep. Var., dependent variable; Treat. Level, treatment level.

Figure 2.5 Layout for a dependent samples *t*-statistic design. Each block in the smoking experiment contains two matched participants. The participants in each block are randomly assigned to the treatment levels. The means of the treatments levels are denoted by $\bar{Y}_{.1}$ and $\bar{Y}_{.2}$.

$$H_0 : \mu_1 - \mu_2 = \delta_0 \qquad (8)$$

$$H_1 : \mu_1 - \mu_2 \neq \delta_0 \qquad (9)$$

where μ_1 and μ_2 denote the means of the respective populations; δ_0 usually is equal to 0. If our hunch is correct, that difficulty in breaking the smoking habit is related to the number of cigarettes smoked per day, the t test for the dependent-samples design should result in a more powerful test of a false null hypothesis than the t test for the independent-samples design. The increased power results from isolating the nuisance variable of number of cigarettes smoked per day and thereby obtaining a more precise estimate of treatment effects and a reduction in the size of the error effects.

In this example, dependent samples were obtained by forming pairs of smokers who were similar with respect to the number of cigarettes smoked per day—a nuisance variable that is positively correlated with the dependent variable. This procedure is called *participant matching*. Dependent samples also can be obtained by: (1) observing each participant under all the treatment levels—that is, obtaining repeated measures on the participants; (2) using identical twins or litter mates in which case the participants have similar genetic characteristics; and (3) obtaining participants who are matched by mutual selection, for example, husband and wife pairs or business partners.

I have described four ways of obtaining dependent samples. The use of repeated measures on the participants usually results in the best within-block homogeneity. However, if repeated measures are obtained, the effects of one treatment level should dissipate before the participant is observed under the other treatment level. Otherwise the second observation will reflect the effects of both treatment levels. There is no such restriction, of course, if carryover effects such as learning or fatigue are the principal interest of the researcher. If blocks are composed of identical twins or litter mates, it is assumed that the performance of participants having identical or similar heredities will be more homogeneous than the performance of participants having dissimilar heredities. If blocks are composed of participants who are matched by mutual selection, for example, husband and wife pairs, a researcher must ascertain that the participants in a block are in fact more homogeneous with respect to the dependent variable than are unmatched participants. A husband and wife often have similar interests and socioeconomic levels; the couple is less likely to have similar mechanical aptitudes.

Solomon four-group design

The Solomon four-group design enables a researcher to control all threats to internal validity as well as some threats to external validity. *External validity* is concerned with the generalizability of research findings to and across populations of participants and settings. The layout for the Solomon four-group design is shown in Figure 2.6.

		Dep. Var.	Treat. Level	Dep. Var.
Group$_1$	Block$_1$	Y_{11}	a_1	Y_{12}
	Block$_2$	Y_{21}	a_1	Y_{22}
	⋮	⋮	⋮	⋮
	Block$_{n_1}$	$Y_{n_1 1}$	a_1	$Y_{n_1 2}$
		$\bar{Y}_{.1}$		$\bar{Y}_{.2}$
Group$_2$	Block$_1$	Y_{13}		Y_{14}
	Block$_2$	Y_{23}		Y_{24}
	⋮	⋮		⋮
	Block$_{n_2}$	$Y_{n_2 3}$		$Y_{n_2 4}$
		$\bar{Y}_{.3}$		$\bar{Y}_{.4}$
Group$_3$	Block$_1$		a_1	Y_{15}
	Block$_2$		a_1	Y_{25}
	⋮		⋮	⋮
	Block$_{n_3}$		a_1	$Y_{n_3 5}$
				$\bar{Y}_{.5}$
Group$_4$	Block$_1$			Y_{16}
	Block$_2$			Y_{26}
	⋮			⋮
	Block$_{n_4}$			$Y_{n_4 6}$
				$\bar{Y}_{.6}$

Figure 2.6 Layout for a Solomon four-group design; *n* participants are randomly assigned to four groups with n_1, n_2, n_3, and n_4 participants in the groups. The pretest is given at the same time to groups 1 and 2. The treatment is administered at the same time to groups 1 and 3. The posttest is given at the same time to all four groups.

The data from the design can be analyzed in a variety of ways. For example, the effects of treatment a_1 can be evaluated using a dependent samples t test of H_0: $\mu_1 - \mu_2$ and independent samples t tests of H_0: $\mu_2 - \mu_4$, H_0: $\mu_3 - \mu_5$, and H_0: $\mu_5 - \mu_6$. Unfortunately, no statistical procedure simultaneously uses all of the data. A two-treatment, completely randomized, factorial design that is described later can be used to evaluate the effects of the treatment, pretesting, and the testing-treatment interaction. The latter effect is a threat to external validity. The layout for the

		Treatment B		
		b_1 No Pretest	b_2 Pretest	
Treatment A	a_1 Treatment	Y_{15} Y_{25} $\vdots$ Y_{n_35}	Y_{12} Y_{22} $\vdots$ Y_{n_12}	
		$\bar{Y}_{.5}$	$\bar{Y}_{.2}$	$\bar{Y}_{\text{Treatment}}$
	a_2 No Treatment	Y_{16} Y_{26} $\vdots$ Y_{n_46}	Y_{14} Y_{24} $\vdots$ Y_{n_24}	
		$\bar{Y}_{.6}$	$\bar{Y}_{.4}$	$\bar{Y}_{\text{No Treatment}}$
		$\bar{Y}_{\text{No Pretest}}$	$\bar{Y}_{\text{Pretest}}$	

Figure 2.7 Layout for analyzing the Solomon four-group design. Three contrasts can be tested: effects of treatment *A* ($\bar{Y}_{\text{Treatment}}$ versus $\bar{Y}_{\text{No Treatment}}$), effects of treatment *B* ($\bar{Y}_{\text{Pretest}}$ versus $\bar{Y}_{\text{No pretest}}$), and the interaction of treatments *A* and *B* ($\bar{Y}_{.5} + \bar{Y}_{.4}$) versus ($\bar{Y}_{.2} + \bar{Y}_{.6}$).

factorial design is shown in Figure 2.7. The null hypotheses are as follows:

$H_0 : \mu_{\text{Treatment}} = \mu_{\text{No Treatment}}$ (treatment A population means are equal)

$H_0 : \mu_{\text{Pretest}} = \mu_{\text{No Pretest}}$ (treatment B population means are equal)

$H_0 : A \times B$ interaction $= 0$ (treatments A and B do not interact)

The null hypotheses are tested with the following F statistics:

$$F = \frac{MSA}{MSWCELL}, F = \frac{MSB}{MSWCELL}, \text{ and}$$
$$F = \frac{MSA \times B}{MSWCELL} \qquad (10)$$

where MSWCELL denotes the within-cell-mean square.

Insight into the effects of administering a pretest is obtained by examining the sample contrasts of $\bar{Y}_{.2}$ versus $\bar{Y}_{.5}$ and $\bar{Y}_{.4}$ versus $\bar{Y}_{.6}$. Similarly, insight into the effects of maturation and history is obtained by examining the sample contrasts of $\bar{Y}_{.1}$ versus $\bar{Y}_{.6}$ and $\bar{Y}_{.3}$ versus $\bar{Y}_{.6}$.

The inclusion of one or more control groups in a design helps to ameliorate the threats to internal validity mentioned earlier. However, as I describe next, a researcher must still deal with other threats to internal validity when the experimental units are people.

THREATS TO INTERNAL VALIDITY WHEN THE EXPERIMENTAL UNITS ARE PEOPLE

Demand characteristics

Doing research with people poses special threats to the internal validity of a design. Because of space restrictions, I will mention only three threats: demand characteristics, participant-predisposition effects, and experimenter-expectancy effects. According to Orne (1962), demand characteristics result from cues in the experimental environment or procedure that lead participants to make inferences about the purpose of an experiment and to respond in accordance with (or in some cases, contrary to) the perceived purpose. People are inveterate problem solvers. When they are told to perform a task, the majority will try to figure out what is expected of them and perform accordingly. Demand characteristics can result from rumors about an experiment, what participants are told when they sign up for an experiment, the laboratory environment, or the communication that occurs during the course of an experiment. Demand characteristics influence a participant's perceptions of what is appropriate or expected and, hence, their behavior.

Participant-predisposition effects

Participant-predisposition effects can also affect the interval validity of an experiment. Because of past experiences, personality, and so on, participants come to experiments with a predisposition to respond in a particular way. I will mention three kinds of participants. The first group of participants is mainly concerned with pleasing the researcher and being 'good subjects.' They try, consciously or unconsciously, to provide data that support the researcher's hypothesis. This participant predisposition is called the *cooperative-participant effect*.

A second group of participants tends to be uncooperative and may even try to sabotage the experiment. Masling (1966) has called this predisposition the '*screw you effect*.' The effect is often seen when research participation is part of a college course requirement. The predisposition can result from resentment over being required to participate in an experiment, from a bad experience in a previous experiment such as being deceived or made to feel inadequate, or from a dislike for the course or the professor associated with the course. Uncooperative participants may try, consciously or unconsciously, to provide data that do not support the researcher's hypothesis.

A third group of participants are apprehensive about being evaluated. Participants with *evaluation apprehension* (Rosenberg, 1965) aren't interested in the experimenter's hypothesis, much less in sabotaging the experiment. Instead, their primary concern is in gaining a positive evaluation from the researcher. The data they provide are colored by a desire to appear intelligent, well adjusted, and so on, and to avoid revealing characteristics that they consider undesirable. Clearly, these participant-predisposition effects can affect the internal validity of an experiment.

Experimenter-expectancy effects

Experimenter-expectancy effects can also affect the internal validity of an experiment. Experiments with human participants are social situations in which one person behaves under the scrutiny of another. The researcher requests a behavior and the participant behaves. The researcher's overt request may be accompanied by other more subtle requests and messages. For example, body language, tone of voice, and facial expressions can communicate the researcher's expectations and desires concerning the outcome of an experiment. Such communications can affect a subject's performance.

Rosenthal (1963) has documented numerous examples of how experimenter-expectancy effects can affect the internal validity of an experiment. He found that researchers tend to obtain from their subjects, whether human or animal, the data they want or expect to obtain. A researcher's expectations and desires also can influence the way he or she records, analyses, and interprets data. According to Rosenthal (1969, 1978), observational or recording errors are usually small and unintentional. However, when such errors occur, more often than not they are in the direction of supporting the researcher's hypothesis. Sheridan (1976) reported that researchers are much more likely to recompute and double-check results that conflict with their hypotheses than results that support their hypotheses.

Kirk (1995: 22–24) described a variety of ways of minimizing these kinds of threats to internal validity. It is common practice, for example, to use a *single-blind procedure* in which participants are not informed about the nature of their treatment and, when feasible, the purpose of the experiment. A single-blind procedure helps to minimize the effects of demand characteristics.

A *double-blind procedure*, in which neither the participant nor the researcher knows which treatment level is administered, is even more effective. For example, in the smoking experiment the patch with the medication and the placebo can be coded so that those administering the treatment cannot identify the condition that is administered. A double-blind procedure helps to minimize both experimenter-expectancy effects and demand characteristics. Often, the nature of the treatment levels is easily identified. In this case, a *partial-blind procedure* can be used in which the researcher does not know until just before administering the treatment level which level will be administered. In this way, experimenter-expectancy effects are minimized up until the administration of the treatment level.

The relatively simple designs described so far provide varying degrees of control of threats to internal validity. For a description of other simple designs such as the longitudinal-overlapping design, time-lag design, time-series design, and single-subject design, the reader is referred to Kirk (1995: 12–16). The designs described next are all analyzed using the analysis of variance and provide excellent control of threats to internal validity.

ANALYSIS OF VARIANCE DESIGNS WITH ONE TREATMENT

Completely randomized design

The simplest analysis of variance (ANOVA) design (a completely randomized design) involves randomly assigning participants to a treatment with two or more levels. The design shares many of the features of an independent-samples t-statistic design. Again, consider the smoking experiment and suppose that the researcher wants to evaluate the effectiveness of three treatment levels: medication delivered by a patch, cognitive-behavioral therapy, and a patch without medication (placebo). In this example and those that follow, a fixed-effects model is assumed, that is, the experiment contains all of the treatment levels of interest to the researcher. The null and alternative hypotheses for the smoking experiment are, respectively:

$$H_0 : \mu_1 = \mu_2 = \mu_3 \qquad (11)$$

$$H_1 : \mu_j \neq \mu_{j'} \text{ for some } j \text{ and } j' \qquad (12)$$

Assume that 30 smokers are available to participate in the experiment. The smokers are randomly assigned to the three treatment levels with 10 smokers assigned to each level. The layout for the experiment is shown in Figure 2.8. Comparison of the layout in this figure with that in Figure 2.4 for an independent samples t-statistic design reveals that they are the same except that the completely randomized design has three treatment levels.

I have identified the null hypothesis that the researcher wants to test, $\mu_1 = \mu_2 = \mu_3$, and described the manner in which the participants are assigned to the three treatment levels. Space limitations prevent me from describing the computational procedures or the assumptions associated with the design. For this information, the reader is referred to the many excellent books on experimental design (Anderson, 2001; Dean and Voss, 1999; Giesbrecht and Gumpertz, 2004; Kirk, 1995; Maxwell and Delaney, 2004; Ryan, 2007).

The partition of the total sum of squares, *SSTOTAL*, and total degrees of freedom, $np - 1$, for a completely randomized design are as follows:

$$SSTOTAL = SSBG + SSWG \qquad (13)$$

$$np - 1 = (p - 1) + p(n - 1) \qquad (14)$$

where *SSBG* denotes the between-groups sum of squares and *SSWG* denotes the within-groups sum of squares. The null hypothesis is tested with the following F statistic:

$$F = \frac{SSBG/(p-1)}{SSWG/[p(n-1)]} = \frac{MSBG}{MSWG} \qquad (15)$$

		Treat. Level	Dep. Var.
	Participant$_1$	a_1	Y_{11}
	Participant$_2$	a_1	Y_{21}
Group$_1$	⋮	⋮	⋮
	Participant$_{n_1}$	a_1	$Y_{n_1 1}$
			$\bar{Y}_{.1}$
	Participant$_1$	a_2	Y_{12}
	Participant$_2$	a_2	Y_{22}
Group$_2$	⋮	⋮	⋮
	Participant$_{n_2}$	a_2	$Y_{n_2 2}$
			$\bar{Y}_{.2}$
	Participant$_1$	a_3	Y_{13}
	Participant$_2$	a_3	Y_{23}
Group$_3$	⋮	⋮	⋮
	Participant$_{n_3}$	a_3	$Y_{n_3 3}$
			$\bar{Y}_{.3}$

Dep. Var., dependent variable.

Figure 2.8 Layout for a completely randomized design with $p = 3$ treatment levels (Treat. Level). Thirty participants are randomly assigned to the treatment levels with $n_1 = n_2 = n_3 = 10$. The means of groups one, two, and three are denoted by, $\bar{Y}_{.1}$, $\bar{Y}_{.2}$, and $\bar{Y}_{.3}$, respectively

The advantages of a completely randomized design are: (1) simplicity in the randomization and statistical analysis; and (2) flexibility with respect to having an equal or unequal number of participants in the treatment levels. A disadvantage is that nuisance variables such as differences among the participants prior to the administration of the treatment are controlled by random assignment. For this control to be effective, the participants should be relatively homogeneous or a relatively large number of participants should be used. The design described next, a randomized block design, enables a researcher to isolate and remove one source of variation among participants that ordinarily would be included in the error effects of the F statistic. As a result, the randomized block design is usually more powerful than the completely randomized design.

Randomized block design

The randomized block design can be thought of as an extension of a dependent samples t-statistic design for the case in which the treatment has two or more levels. The layout for a randomized block design with $p = 3$ levels of treatment A and $n = 10$ blocks is shown in Figure 2.9. Comparison of the layout in this figure with that in Figure 2.5 for a dependent samples t-statistic design reveals that they are the same except that the randomized block design has three treatment levels. A block can contain a single participant who is observed under all p treatment levels or p participants who are similar with respect to a variable that is positively correlated with the dependent variable. If each block contains one participant, the order in which the treatment levels are administered is randomized independently for each block, assuming that the nature of the treatment and the research hypothesis permit this. If a block contains p matched participants, the participants in each block are randomly assigned to the treatment levels. The use of repeated measures or matched participants does not affect the statistical analysis. However, the alternative procedures do affect the interpretation of the results. For example, the results of an experiment with repeated measures generalize to a population of participants who have been exposed to all of the treatment levels. The results of an experiment with matched participants generalize to a population of participants who have been exposed to only one treatment level. Some writers reserve the designation *randomized block design* for this latter case. They refer to a design with repeated measurements in which the order of administration of the treatment levels is randomized independently for each participant as a *subjects-by-treatments design*. A design with repeated measurements in which the order of administration of the treatment levels is the same for all participants is referred to as a *subject-by-trials* design.

	Treat. Level	Dep. Var.	Treat. Level	Dep. Var.	Treat. Level	Dep. Var.	
Block$_1$	a_1	Y_{11}	a_2	Y_{12}	a_3	Y_{13}	$\bar{Y}_{1.}$
Block$_2$	a_1	Y_{21}	a_2	Y_{22}	a_3	Y_{23}	$\bar{Y}_{2.}$
Block$_3$	a_1	Y_{31}	a_2	Y_{32}	a_3	Y_{33}	$\bar{Y}_{3.}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Block$_{10}$	a_1	$Y_{10,1}$	a_2	$Y_{10,2}$	a_3	$Y_{10,3}$	$\bar{Y}_{10.}$
		$\bar{Y}_{.1}$		$\bar{Y}_{.2}$		$\bar{Y}_{.3}$	

Dep. Var., dependent variable.

Figure 2.9 Layout for a randomized block design with $p = 3$ treatment levels (Treat. Level) and $n = 10$ blocks. In the smoking experiment, the p participants in each block were randomly assigned to the treatment levels. The means of treatment A are denoted by $\bar{Y}_{.1}$, $\bar{Y}_{.2}$, and $\bar{Y}_{.3}$; the means of the blocks are denoted by $\bar{Y}_{1.}, \ldots, \bar{Y}_{10.}$.

I prefer to use the designation *randomized block design* for all three cases.

The total sum of squares and total degrees of freedom are partitioned as follows:

$$SSTOTAL = SSA + SSBLOCKS + SSRESIDUAL \quad (16)$$

$$np - 1 = (p-1) + (n-1) + (n-1)(p-1) \quad (17)$$

where *SSA* denotes the treatment *A* sum of squares and *SSBLOCKS* denotes the block sum of squares. The *SSRESIDUAL* is the interaction between treatment *A* and blocks; it is used to estimate error effects. Two null hypotheses can be tested:

$$H_0 : \mu_{.1} = \mu_{.2} = \mu_{.3} \quad \text{(treatment } A \text{ population means are equal)} \quad (18)$$

$$H_0 : \mu_{1.} = \mu_{2.} = \ldots = \mu_{.15} \quad \text{(block, } BL\text{, population means are equal)} \quad (19)$$

where μ_{ij} denotes the population mean for the ith block and the jth level of treatment A. The F statistics are:

$$F = \frac{SSA/(p-1)}{SSRESIDUAL/[(n-1)(p-1)]} = \frac{MSA}{MSRESIDUAL} \quad (20)$$

$$F = \frac{SSBL/(n-1)}{SSRESIDUAL/[(n-1)(p-1)]} = \frac{MSBL}{MSRESIDUAL} \quad (21)$$

The test of the block null hypothesis is of little interest because the blocks represent a nuisance variable that a researcher wants to isolate so that it does not appear in the estimators of the error effects.

The advantages of this design are: (1) simplicity in the statistical analysis; and (2) the ability to isolate a nuisance variable so as to obtain greater power to reject a false null hypothesis. The disadvantages of the design include: (1) the difficulty of forming homogeneous blocks or observing participants p times when p is large; and (2) the restrictive assumptions (sphericity and additivity) of the design. For a description of these assumptions, see Kirk (1995: 271–282).

Latin square design

The Latin square design is appropriate to experiments that have one treatment, denoted by A, with $p \geq 2$ levels and two nuisance variables, denoted by B and C, each with p levels. The design gets its name from an ancient puzzle that was concerned with the number of ways that Latin letters can be arranged in a square matrix so that each letter appears once in each row and once in each column. A 3 × 3 Latin square is shown in Figure 2.10.

The randomized block design enables a researcher to isolate one nuisance variable: variation among blocks. A Latin square design extends this procedure to two nuisance

	c_1	c_2	c_3
b_1	a_1	a_2	a_3
b_2	a_2	a_3	a_1
b_3	a_3	a_1	a_2

Figure 2.10 Three-by-three Latin square, where a_j denotes one of the $j = 1, \ldots, p$ levels of treatment A, b_k denotes one of the $k = 1, \ldots, p$ levels of nuisance variable B, and c_l denotes one of the $l = 1, \ldots, p$ levels of nuisance variable C. Each level of treatment A appears once in each row and once in each column as required for a Latin square.

variables: variation associated with the rows (B) of the Latin square and variation associated with the columns of the square (C). As a result, the Latin square design is generally more powerful than the randomized block design. The layout for a Latin square design with three levels of treatment A is shown in Figure 2.11 and is based on the $a_jb_kc_l$ combinations in Figure 2.10. The total sum of squares and total degrees of freedom are partitioned as follows:

$$\begin{aligned} SSTOTAL &= SSA + SSB + SSC \\ &\quad + SSRESIDUAL + SSWCELL \\ np^2 - 1 &= (p-1) + (p-1) + (p-1) \\ &\quad + (p-1)(p-2) + p^2(n-1) \end{aligned}$$

where SSA denotes the treatment sum of squares, SSB denotes the row sum of squares, and SSC denotes the column sum of squares. $SSWCELL$ denotes the within cell sum of squares and estimates error effects. Four null hypotheses can be tested:

$H_0 : \mu_{1..} = \mu_{2..} = \mu_{3..}$ (treatment A population means are equal)

$H_0 : \mu_{.1.} = \mu_{.2.} = \mu_{.3.}$ (row, B, population means are equal)

$H_0 : \mu_{..1} = \mu_{..2} = \mu_{..3}$ (column, C, population means are equal)

H_0 : interaction components = 0 (selected $A \times B, A \times C, B \times C,$ and $A \times B \times C$ interaction components equal zero)

		Treat. Comb.	Dep. Var.
Group$_1$	Participant$_1$	$a_1b_1c_1$	Y_{111}
	⋮	⋮	⋮
	Participant$_{n_1}$	$a_1b_1c_1$	Y_{111}
			$\bar{Y}_{.111}$
Group$_2$	Participant$_1$	$a_1b_2c_3$	Y_{123}
	⋮	⋮	⋮
	Participant$_{n_2}$	$a_1b_2c_3$	Y_{123}
			$\bar{Y}_{.123}$
Group$_3$	Participant$_1$	$a_1b_3c_2$	Y_{132}
	⋮	⋮	⋮
	Participant$_{n_3}$	$a_1b_3c_2$	Y_{132}
			$\bar{Y}_{.132}$
Group$_4$	Participant$_1$	$a_2b_1c_2$	Y_{212}
	⋮	⋮	⋮
	Participant$_{n_4}$	$a_2b_1c_2$	Y_{212}
			$\bar{Y}_{.212}$
	⋮	⋮	⋮
Group$_9$	Participant$_1$	$a_3b_3c_1$	Y_{331}
	⋮	⋮	⋮
	Participant$_{n_9}$	$a_3b_3c_1$	Y_{331}
			$\bar{Y}_{.331}$

Dep. Var., dependent variable; Treat. Comb., treatment combination.

Figure 2.11 Layout for a Latin square design that is based on the Latin square in Figure 2.10. Treatment A and the two nuisance variables, B and C, each have $p = 3$ levels.

where μ_{jkl} denotes a population mean for the jth treatment level, kth row, and lth column. The F statistics are:

$$F = \frac{SSA/(p-1)}{SSWCELL/[p^2(n-1)]} = \frac{MSA}{MSWCELL}$$

$$F = \frac{SSB/(p-1)}{SSWCELL/[p^2(n-1)]} = \frac{MSB}{MSWCELL}$$

$$F = \frac{SSC/(p-1)}{SSWCELL/[p^2(n-1)]} = \frac{MSC}{MSWCELL}$$

$$F = \frac{SSRESIDUAL/(p-1)(p-2)}{SSWCELL/[p^2(n-1)]} = \frac{MSRESIDUAL}{MSWCELL}.$$

The advantage of the Latin square design is the ability to isolate two nuisance variables to obtain greater power to reject a false null hypothesis. The disadvantages are: (1) the number of treatment levels, rows, and columns of the Latin square must be equal, a balance that may be difficult to achieve; (2) if there are any interactions among the treatment levels, rows, and columns, the test of treatment A is positively biased; and (3) the randomization is relatively complex.

I have described three simple ANOVA designs: completely randomized design, randomized block design, and Latin square design. I call these three designs *building-block designs* because all complex ANOVA designs can be constructed by modifying or combining these simple designs (Kirk, 2005: 69). Furthermore, the randomization procedures, data-analysis procedures, and assumptions for complex ANOVA designs are extensions of those for the three building-block designs. The generalized randomized block design that is described next represents a modification of the randomized block design.

Generalized randomized block design

A generalized randomized block design is a variation of a randomized block design. Instead of having n blocks of homogeneous participants, the generalized randomized block design has w groups of np homogeneous participants. The $z = 1, \ldots, w$ groups, like the blocks in a randomized block design, represent a nuisance variable that a researcher wants to remove from the error effects. The generalized randomized block design is appropriate for experiments that have one treatment with $p \geq 2$ treatment levels and w groups each containing np homogeneous participants. The total number of participants in the design is $N = npw$. The np participants in each group are randomly assigned to the p treatment levels with the restriction that n participants are assigned to each level. The layout for the design is shown in Figure 2.12.

In the smoking experiment, suppose that 30 smokers are available to participate. The 30 smokers are ranked with respect to the length of time that they have smoked. The $np = (2)(3) = 6$ smokers who have smoked for the shortest length of time are assigned to group 1, the next six smokers are assigned to group 2, and so on. The six smokers in each group are then randomly assigned to the three treatment levels with the restriction that $n = 2$ smokers are assigned to each level.

The total sum of squares and total degrees of freedom are partitioned as follows:

$$\begin{aligned} SSTOTAL &= SSA + SSG \\ &\quad + SSA \times G + SSWCELL \\ npw - 1 &= (p-1) + (w-1) \\ &\quad + (p-1)(w-1) \\ &\quad + pw(n-1) \end{aligned}$$

where SSG denotes the groups sum of squares and $SSA \times G$ denotes the interaction of treatment A and groups. The within cells sum of squares, $SSWCELL$, is used to estimate error effects. Three null hypotheses can be tested:

$H_0 : \mu_{1.} = \mu_{2.} = \mu_{3.}$ (treatment A population means are equal)

$H_0 : \mu_{.1} = \mu_{.2} = \cdots = \mu_{.5}$ (group population means are equal)

$H_0 : A \times G$ interaction $= 0$ (treatment A and groups do not interact),

where μ_{jz} denotes a population mean for the ith block, jth treatment level, and zth group. The three null hypotheses are tested using the following F statistics:

$$F = \frac{SSA/(p-1)}{SSWCELL/[pw(n-1)]} = \frac{MSA}{MSWCELL}$$

		Treat. Level	Dep. Var.		Treat. Level	Dep. Var.		Treat. Level	Dep. Var.
Group_1	1	a_1	Y_{111}	3	a_2	Y_{321}	5	a_3	Y_{531}
	2	a_1	Y_{211}	4	a_2	Y_{421}	6	a_3	Y_{631}
			$\bar{Y}_{.11}$			$\bar{Y}_{.21}$			$\bar{Y}_{.31}$
Group_2	7	a_1	Y_{712}	9	a_2	Y_{922}	11	a_3	$Y_{11,32}$
	8	a_1	Y_{812}	10	a_2	$Y_{10,22}$	12	a_3	$Y_{12,32}$
			$\bar{Y}_{.12}$			$\bar{Y}_{.22}$			$\bar{Y}_{.32}$
Group_3	13	a_1	$Y_{13,13}$	15	a_2	$Y_{15,23}$	17	a_3	$Y_{17,33}$
	14	a_1	$Y_{14,13}$	16	a_2	$Y_{16,23}$	18	a_3	$Y_{18,33}$
			$\bar{Y}_{.13}$			$\bar{Y}_{.23}$			$\bar{Y}_{.33}$
Group_4	19	a_1	$Y_{19,14}$	21	a_2	$Y_{21,24}$	23	a_3	$Y_{23,34}$
	20	a_1	$Y_{20,14}$	22	a_2	$Y_{22,24}$	24	a_3	$Y_{24,34}$
			$\bar{Y}_{.14}$			$\bar{Y}_{.24}$			$\bar{Y}_{.34}$
Group_5	25	a_1	$Y_{25,15}$	27	a_2	$Y_{27,25}$	29	a_3	$Y_{29,35}$
	26	a_1	$Y_{26,15}$	28	a_2	$Y_{28,25}$	30	a_3	$Y_{30,35}$
			$\bar{Y}_{.15}$			$\bar{Y}_{.25}$			$\bar{Y}_{.35}$

Dep. Var., dependent variable.

Figure 2.12 Generalized randomized block design with $n = 30$ participants, $p = 3$ treatment levels, and $w = 5$ groups of $np = (2)(3) = 6$ homogeneous participants. The six participants in each group are randomly assigned to the three treatment levels (Treat. Level) with the restriction that two participants are assigned to each level.

$$F = \frac{SSG/(w-1)}{SSWCELL/[pw(n-1)]} = \frac{MSG}{MSWCELL}$$

$$F = \frac{SSA \times G/(p-1)(w-1)}{SSWCELL/[pw(n-1)]} = \frac{MSA \times G}{MSWCELL}$$

The generalized randomized block design enables a researcher to isolate one nuisance variable, an advantage that it shares with the randomized block design. Furthermore, the design uses the within cell variation in the $pw = 15$ cells to estimate error effects rather than an interaction as in the randomized block design. Hence, the restrictive sphericity and additivity assumptions of the randomized block design are replaced with the assumption of homogeneity of within cell population variances.

ANALYSIS OF VARIANCE DESIGNS WITH TWO OR MORE TREATMENTS

The ANOVA designs described thus far all have one treatment with $p \geq 2$ levels. The designs described next have two or more treatments denoted by the letters A, B, C, and so on. If all of the treatments are completely crossed, the design is called a *factorial design*. Two treatments are completely crossed if each level of one treatment appears in combination with each level of the other treatment and vice se versa. Alternatively, a treatment can be nested within another treatment. If, for example, each level of treatment B appears with only one level of treatment A, treatment B is nested within treatment A. The distinction

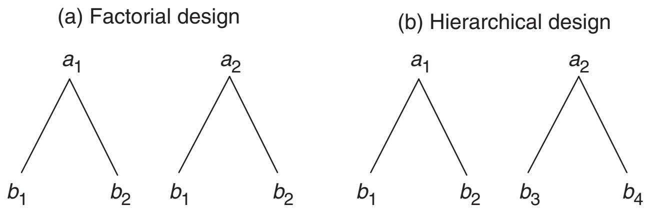


Figure 2.13 (a) illustrates crossed treatments. In (b), treatment *B*(*A*) is nested in treatment *A*.

between crossed and nested treatments is illustrated in Figure 2.13. The use of crossed treatments is a distinguishing characteristic of all factorial designs. The use of at least one nested treatment in a design is a distinguishing characteristic of *hierarchical designs*.

Completely randomized factorial design

The simplest factorial design from the standpoint of randomization procedures and data analysis is the *completely randomized factorial design* with p levels of treatment A and q levels of treatment B. The design is constructed by crossing the p levels of one completely randomized design with the q levels of a second completely randomized design. The design has $p \times q$ treatment combinations, $a_1b_1, a_1b_2, \ldots, a_pb_q$.

The layout for a two-treatment completely randomized factorial design with $p = 2$ levels of treatment A and $q = 2$ levels of treatment B is shown in Figure 2.14. In this example, 40 participants are randomly assigned to the $2 \times 2 = 4$ treatment combinations with the restriction that $n = 10$ participants are assigned to each combination. This design enables a researcher to simultaneously evaluate two treatments, A and B, and, in addition, the interaction between the treatments, denoted by $A \times B$. Two treatments are said to interact if differences in the dependent variable for the levels of one treatment are different at two or more levels of the other treatment.

		Treat. Comb.	Dep. Var.
Group$_1$	Participant$_1$	a_1b_1	Y_{111}
	⋮	⋮	⋮
	Participant$_{n_1}$	a_1b_1	Y_{n_111}
			$\bar{Y}_{.11}$
Group$_2$	Participant$_1$	a_1b_2	Y_{112}
	⋮	⋮	⋮
	Participant$_{n_2}$	a_1b_2	Y_{n_212}
			$\bar{Y}_{.12}$
Group$_3$	Participant$_1$	a_2b_1	Y_{121}
	⋮	⋮	⋮
	Participant$_{n_3}$	a_2b_1	Y_{n_321}
			$\bar{Y}_{.21}$
Group$_4$	Participant$_1$	a_2b_2	Y_{122}
	⋮	⋮	⋮
	Participant$_{n_4}$	a_2b_2	Y_{n_422}
			$\bar{Y}_{.22}$

Dep. Var., dependent variable; Treat. Comb., treatment combination.

Figure 2.14 Layout for a two-treatment, completely randomized factorial design where $n = 40$ participants were randomly assigned to four combinations of treatments *A* and *B*, with the restriction that $n_1 = \ldots = n_4 = 10$ participants were assigned to each combination.

The total sum of squares and total degrees of freedom for the design are partitioned as follows:

$$\begin{aligned} SSTOTAL &= SSA + SSB \\ &\quad + SSA \times B + SSWCELL \\ npq - 1 &= (p-1) + (q-1) \\ &\quad + (p-1)(q-1) + pq(n-1), \end{aligned}$$

where $SSA \times B$ denotes the interaction sum of squares for treatments A and B. Three null hypotheses can be tested:

$H_0: \mu_{1.} = \mu_{2.} = \cdots = \mu_{p.}$ (treatment A population means are equal)

$H_0: \mu_{.1} = \mu_{.2} = \cdots = \mu_{.q}$ (treatment B population means are equal)

$H_0: A \times B$ interaction $= 0$ (treatments A and B do not interact),

where μ_{jk} denotes a population mean for the jth level of treatment A and the kth level of treatment B. The F statistics for testing the null hypotheses are as follows:

$$F = \frac{SSA/(p-1)}{SSWCELL/[pq(n-1)]} = \frac{MSA}{MSWCELL}$$
$$F = \frac{SSB/(q-1)}{SSWCELL/[pq(n-1)]} = \frac{MSB}{MSWCELL}$$
$$F = \frac{SSA \times B/(p-1)(q-1)}{SSWCELL/[pq(n-1)]} = \frac{MSA \times B}{MSWCELL}.$$

The advantages of a completely randomized factorial design are as follows: (1) All participants are used in simultaneously evaluating the effects of two or more treatments. The effects of each treatment are evaluated with the same precision as if the entire experiment had been devoted to that treatment alone. Thus, the design permits efficient use of resources. (2) A researcher can determine whether the treatments interact. The disadvantages of the design are as follows: (1) If numerous treatments are included in the experiment, the number of participants required can be prohibitive. (2) A factorial design lacks simplicity in the interpretation of results if interaction effects are present. Unfortunately, interactions among variables in the behavioral sciences and education are common. (3) The use of a factorial design commits a researcher to a relatively large experiment. A series of small experiments permit greater freedom in pursuiting unanticipated promising lines of investigation.

Randomized block factorial design

A two-treatment, *randomized block factorial design* is constructed by crossing p levels of one randomized block design with q levels of a second randomized block design. This procedure produces $p \times q$ treatment combinations: $a_1b_1, a_1b_2, \ldots, a_pb_q$. The design uses the blocking technique described in connection with a randomized block design to isolate variation attributable to a nuisance variable while simultaneously evaluating two or more treatments and associated interactions.

A two-treatment, randomized block factorial design has blocks of size pq. If a block consists of matched participants, n blocks of pq matched participants must be formed. The participants in each block are randomly assigned to the pq treatment combinations. Alternatively, if repeated measures are obtained, each participant is observed pq times. For this case, the order in which the treatment combinations is administered is randomized independently for each block, assuming that the nature of the treatments and research hypotheses permit this. The layout for the design with $p = 2$ and $q = 2$ treatment levels is shown in Figure 2.15.

The total sum of squares and total degrees of freedom for a randomized block factorial design are partitioned as follows:

$$SSTOTAL = SSBL + SSA + SSB + SSA \times B + SSRESIDUAL$$
$$npq - 1 = (n-1) + (p-1) + (q-1) + (p-1)(q-1) + (n-1)(pq-1).$$

Four null hypotheses can be tested:

$H_0: \mu_{1..} = \mu_{2..} = \ldots = \mu_{n..}$ (block population means are equal)

$H_0: \mu_{.1.} = \mu_{.2.} = \ldots = \mu_{.p.}$ (treatment A population means are equal)

$H_0: \mu_{..1} = \mu_{..2} = \ldots = \mu_{..q}$ (treatment B population means are equal)

$H_0: A \times B$ interaction $= 0$ (treatments A and B do not interact),

where μ_{ijk} denotes a population mean for the ith block, jth level of treatment A, and kth level

	Treat. Comb.	Dep. Var.	Treat. Comb.	Dep. Var.	Treat. Comb.	Dep. Var.	Treat. Comb.	Dep. Var.	
Block$_1$	a_1b_1	Y_{111}	a_1b_2	Y_{112}	a_2b_1	Y_{121}	a_2b_2	Y_{122}	$\bar{Y}_{1..}$
Block$_2$	a_1b_1	Y_{211}	a_1b_2	Y_{212}	a_2b_1	Y_{221}	a_2b_2	Y_{222}	$\bar{Y}_{2..}$
Block$_3$	a_1b_1	Y_{311}	a_1b_2	Y_{312}	a_2b_1	Y_{321}	a_2b_2	Y_{322}	$\bar{Y}_{3..}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Block$_{10}$	a_1b_1	$Y_{10,11}$	a_1b_2	$Y_{10,12}$	a_2b_1	$Y_{10,21}$	a_2b_2	$Y_{10,22}$	$\bar{Y}_{10,..}$
		$\bar{Y}_{.11}$		$\bar{Y}_{.12}$		$\bar{Y}_{.21}$		$\bar{Y}_{.22}$	

Dep. Var., dependent variable; Treat. Comb., treatment combination.

Figure 2.15 Layout for a two-treatment, randomized block factorial design where four matched participants are randomly assigned to the *pq* = 2 × 2 = 4 treatments combinations in each block.

of treatment *B*. The *F* statistics for testing the null hypotheses are as follows:

$$\begin{aligned}
F &= \frac{SSBL/(n-1)}{SSRESIDUAL/[(n-1)(\,pq-1)]} \\
&= \frac{MSBL}{MSRESIDUAL} \\
F &= \frac{SSA/(\,p-1)}{SSRESIDUAL/[(n-1)(\,pq-1)]} \\
&= \frac{MSA}{MSRESIDUAL} \\
F &= \frac{SSB/(q-1)}{SSRESIDUAL/[(n-1)(\,pq-1)]} \\
&= \frac{MSB}{MSRESIDUAL} \\
F &= \frac{SSA \times B/(\,p-1)(q-1)}{SSRESIDUAL/[(n-1)(\,pq-1)]} \\
&= \frac{MSA \times B}{MSRESIDUAL}.
\end{aligned}$$

The design shares the advantages and disadvantages of the randomized block design. It has an additional disadvantage: if treatment *A* or *B* has numerous levels, say four or five, the block size becomes prohibitively large. Designs that reduce the size of the blocks are described next.

ANALYSIS OF VARIANCE DESIGNS WITH CONFOUNDING

An important advantage of a randomized block factorial design relative to a completely randomized factorial design is superior power. However, if either *p* or *q* in a two-treatment, randomized factorial design is moderately large, the number of treatment combinations in each block can be prohibitively large. For example, if $p = 3$ and $q = 4$, the design has blocks of size $3 \times 4 = 12$. Obtaining *n* blocks with twelve matched participants or observing *n* participants on twelve occasions is generally not feasible. In the late 1920s, Ronald A. Fisher and Frank Yates addressed the problem of prohibitively large block sizes by developing confounding schemes in which only a portion of the treatment combinations in an experiment are assigned to each block (Yates, 1937). Their work was extended in the 1940s by David J. Finney (1945, 1946) and Oscar Kempthorne (1947).

The split-plot factorial design that is described next achieves a reduction in the block size by confounding one or more treatments with groups of blocks. *Group-treatment confounding* occurs when the effects of, say, treatment *A* with *p* levels are indistinguishable from the effects of *p* groups of blocks. This form of confounding is characteristic of all split-plot factorial designs. A second form of confounding, *group-interaction confounding*, also reduces the size of blocks. This form of confounding is characteristic of all confounded factorial designs. A third form of confounding, *treatment-interaction confounding* reduces the number of treatment combinations that must be included in a design. Treatment-interaction confounding is characteristic of all fractional factorial designs.

Reducing the size of blocks and the number of treatment combinations that must be included in a design are attractive. However, in the design of experiments, researchers do not get something for nothing. In the case of a split-plot factorial design, the effects of the confounded treatment are evaluated with less precision than in a randomized block factorial design.

Split-plot factorial design

The *split-plot factorial* design is appropriate for experiments with two or more treatments where the number of treatment combinations exceeds the desired block size. The term *split-plot* comes from agricultural experimentation where the levels of, say, treatment A are applied to relatively large plots of land—the whole plots. The whole plots are then split or subdivided, and the levels of treatment B are applied to the subplots within each whole plot.

A two-treatment, split-plot factorial design is constructed by combining features of two different building-block designs; a completely randomized design with p levels and a randomized block design with q levels. The layout for a split-plot factorial design with treatment A as the between-block treatment and treatment B as the within-blocks treatment is shown in Figure 2.16.

Comparison of Figure 2.16 for the split-plot factorial design with Figure 2.15 for a randomized block factorial design reveals that the designs contain the same treatment combinations: a_1b_1, a_1b_2, a_2b_1, and a_2b_2. However, the block size in the split-plot factorial design is half as large. The smaller block size is achieved by confounding two groups of blocks with treatment A. Consider the sample means for treatment A in Figure 2.16. The difference between $\bar{Y}_{.1.}$ and $\bar{Y}_{.2.}$ reflects the difference between the two groups of blocks as well as the difference between the two levels of treatment A. To put it another way, you cannot tell how much of the difference between $\bar{Y}_{.1.}$ and $\bar{Y}_{.2.}$ is attributable to the difference between Group$_1$ and Group$_2$ and how much is attributable to the difference between treatment levels a_1 and a_2. For this reason, the groups of blocks and treatment A are said to be confounded.

The total sum of squares and total degrees of freedom for a split-plot factorial design are partitioned as follows:

$$\begin{aligned} SSTOTAL &= SSA + SSBL(A) + SSB \\ &\quad + SSA \times B + SSRESIDUAL \\ npq - 1 &= (p-1) + p(n-1) + (q-1) \\ &\quad + (p-1)(q-1) + p(n-1)(q-1), \end{aligned}$$

			Treat. Comb.	Dep. Var.	Treat. Comb.	Dep. Var.	
a_1	Group$_1$	Block$_1$	a_1b_1	Y_{111}	a_1b_1	Y_{112}	$\bar{Y}_{.1.}$
		Block$_2$	a_1b_1	Y_{211}	a_1b_2	Y_{212}	
		$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
		Block$_{n_1}$	a_1b_1	Y_{n_111}	a_1b_2	Y_{n_112}	
a_2	Group$_2$	Block$_1$	a_2b_1	Y_{121}	a_2b_2	Y_{122}	$\bar{Y}_{.2.}$
		Block$_2$	a_2b_1	Y_{221}	a_2b_2	Y_{222}	
		$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
		Block$_{n_2}$	a_2b_1	Y_{n_221}	a_2b_2	Y_{n_222}	
				$\bar{Y}_{..1}$		$\bar{Y}_{..2}$	

Dep. Var., dependent variable; Treat. Comb., treatment combination.

Figure 2.16 Layout for a two-treatment, split-plot factorial design. Treatment *A*, a between-blocks treatment, is confounded with groups. Treatment *B* is a within-blocks treatment.

where *SSBL(A)* denotes the sum of squares for blocks within treatment *A*. Three null hypotheses can be tested:

$H_0 : \mu_{1.} = \mu_{2.} = \ldots = \mu_{p.}$ (treatment *A* population means are equal)

$H_0 : \mu_{.1} = \mu_{.2} = \ldots = \mu_{.q}$ (treatment *B* population means are equal)

$H_0 : A \times B$ interaction $= 0$ (treatments *A* and *B* do not interact),

where μ_{ijk} denotes the *i*th block, *j*th level of treatment *A*, and *k*th level of treatment *B*. The *F* statistics are:

$$F = \frac{SSA/(p-1)}{SSBL(A)/[p(n-1)]} = \frac{MSA}{MSBL(A)}$$

$$F = \frac{SSB/(q-1)}{SSRESIDUAL/[p(n-1)(q-1)]} = \frac{MSB}{MSRESIDUAL}$$

$$F = \frac{SSA \times B/(p-1)(q-1)}{SSRESIDUAL/[p(n-1)(q-1)]} = \frac{MSA \times B}{MSRESIDUAL}$$

Notice that the split-plot factorial design uses two error terms: *MSBL(A)* is used to test treatment *A*; a different and usually much smaller error term, *MSRESIDUAL*, is used to test treatment *B* and the *A* × *B* interaction. The statistic $F = MSA/MSBL(A)$ is like the *F* statistic in a completely randomized design. Similarly, the statistic $F = MSB/MSRESIDUAL$ is like the *F* statistic for treatment *B* in a randomized block design. Because *MSRESIDUAL* is generally smaller than *MSBL(A)*, the power of the tests of treatment *B* and the *A* × *B* interaction is greater than that for treatment *A*. A randomized block factorial design, on the other hand, uses *MSRESIDUAL* to test all three null hypotheses. As a result, the power of the test of treatment *A*, for example, is the same as that for treatment *B*. When tests of treatments *A* and *B* and the *A* × *B* interaction are of equal interest, a randomized block factorial design is a better design choice than a split-plot factorial design. However, if the larger block size is not acceptable and a researcher is more interested in treatment *B* and the *A* × *B* interaction than in treatment *A*, a split-plot factorial design is a good design choice.

Alternatively, if a large block size is not acceptable and the researcher is primarily interested in treatments *A* and *B*, a confounded-factorial design is a good choice. This design, which is described next, achieves a reduction in block size by confounding groups of blocks with the *A* × *B* interaction. As a result, tests of treatments *A* and *B* are more powerful than the test of the *A* × *B* interaction.

Confounded factorial designs

Confounded factorial designs are constructed from either randomized block designs or Latin square designs. One of the simplest completely confounded factorial designs is constructed from two randomized block designs. The layout for the design with $p = q = 2$ is shown in Figure 2.17. The design confounds the *A* × *B* interaction with groups of blocks and thereby reduces the block size to two. The power of the tests of the two treatments is usually much greater than the power of the test of the *A* × *B* interaction. Hence, the completely confounded factorial design is a good design choice if a small block size is required and the researcher is primarily interested in tests of treatments *A* and *B*.

Fractional factorial design

Confounded factorial designs reduce the number of treatment combinations that appear in each block. Fractional-factorial designs use treatment-interaction confounding to reduce the number of treatment combinations that appear in an experiment. For example, the number of treatment combinations in an experiment can be reduced to some fraction—$^1/_2$, $^1/_3$, $^1/_4$, $^1/_8$, $^1/_9$, and so on—of the total number of treatment combinations in an unconfounded factorial design. Unfortunately, a researcher pays a price for using only a fraction of the treatment combinations: ambiguity in interpreting the outcome of the experiment. Each sum of squares has two or

			Treat. Comb.	Dep. Var.	Treat. Comb.	Dep. Var.
		Block$_1$	a_1b_1	Y_{111}	a_2b_2	Y_{122}
		Block$_2$	a_1b_1	Y_{211}	a_2b_2	Y_{222}
a_jb_k	Group$_1$	⋮	⋮	⋮	⋮	⋮
		Block$_{n_1}$	a_1b_1	Y_{n_111}	a_2b_2	Y_{n_122}
				$\bar{Y}_{.11}$		$\bar{Y}_{.22}$
		Block$_1$	a_1b_2	Y_{112}	a_2b_1	Y_{121}
		Block$_2$	a_1b_2	Y_{212}	a_2b_1	Y_{221}
a_jb_k	Group$_2$	⋮	⋮	⋮	⋮	⋮
		Block$_{n_2}$	a_1b_2	Y_{n_212}	a_2b_1	Y_{n_221}
				$\bar{Y}_{.12}$		$\bar{Y}_{.21}$

Dep. Var., dependent variable; Treat. Comb., treatment combination.

Figure 2.17 Layout for a two-treatment, randomized block, confounded factorial design. The *A* × *B* interaction is confounded with groups. Treatments *A* and *B* are within-blocks treatments.

more labels called *aliases*. For example, two labels for the same sum of squares might be treatment *A* and the $B \times C \times D$ interaction.

You may wonder why anyone would use such a design—after all, experiments are supposed to help us resolve ambiguity not create it. Fractional factorial designs are typically used in exploratory research situations where a researcher is interested in six-or- more treatments and can perform follow-up experiments if necessary. The design enables a researcher to efficiently investigate a large number of treatments in an initial experiment, with subsequent experiments designed to focus on the most promising lines of investigation or to clarify the interpretation of the original analysis. The designs are used infrequently in the behavioral sciences and education.

HIERARCHICAL DESIGNS

The multitreatment designs that I have discussed up to now all have crossed treatments. Often researchers in the behavioral sciences and education design experiments in which one of more treatments is nested. Treatment *B* is *nested* in treatment *A* if each level of treatment *B* appears with only one level of treatment *A*. A *hierarchical design* has at least one nested treatment; the remaining treatments are either nested or crossed.

Hierarchical designs are constructed from two or more or a combination of completely randomized and randomized block designs. For example, a researcher who is interested in the effects two types of radiation on learning in rats could assign 30 rats randomly to six cages. The six cages are then randomly assigned to the two types of radiation. The cages restrict the movement of the rats and insure that all rats in a cage receive the same amount of radiation. In this example, the cages are nested in the two types of radiation. The layout for the design is shown in Figure 2.18.

There is a wide array of hierarchical designs. For a description of these designs, the reader is referred to the extensive treatment in Chapter 11 of Kirk (1995).

ANALYSIS OF COVARIANCE

The discussion so far has focused on designs that use *experimental control* to reduce error variance and minimize the effects of nuisance

		Treat. Comb.	Dep. Var.
Group$_1$	Animal$_1$	$a_1 b_1$	Y_{111}
	⋮	⋮	⋮
	Animal$_{n_1}$	$a_1 b_1$	$Y_{n_1 11}$
			$\bar{Y}_{.11}$
Group$_2$	Animal$_1$	$a_1 b_2$	Y_{112}
	⋮	⋮	⋮
	Animal$_{n_2}$	$a_1 b_2$	$Y_{n_2 12}$
			$\bar{Y}_{.12}$
Group$_3$	Animal$_1$	$a_1 b_3$	Y_{113}
	⋮	⋮	⋮
	Animal$_{n_3}$	$a_1 b_3$	$Y_{n_3 13}$
			$\bar{Y}_{.21}$
	⋮	⋮	⋮
Group$_{10}$	Animal$_1$	$a_2 b_{10}$	$Y_{12,10}$
	⋮	⋮	⋮
	Animal$_{n_{10}}$	$a_2 b_{10}$	$Y_{n_{10} 2,10}$
			$\bar{Y}_{.22}$

Dep. Var., dependent variable; Treat. Comb., treatment combination.

Figure 2.18 Layout for a two-treatment, completely randomized, hierarchical design in which $q = 10$ levels of treatment $B(A)$ are nested in $p = 2$ levels of treatment A. Fifty animals are randomly assigned to ten combinations of treatments A and $B(A)$ with the restriction that five animals are assigned to each treatment combination.

variables. Experimental control can take different forms such as random assignment of participants to treatment levels, stratification of participants into homogeneous blocks, and refinement of techniques for measuring a dependent variable. *Analysis of covariance* is an alternative approach to reducing error variance and minimizing the effects of nuisance variables. The approach combines regression analysis with ANOVA and involves measuring one or more *concomitant variables*, also called *covariates*, in addition to the dependent variable. The concomitant variable represents a source of variation that was not controlled in the experiment and a source that is believed to affect the dependent variable.

Analysis of covariance enables a researcher to: (1) remove that portion of the dependent-variable error variance that is predictable from a knowledge of the concomitant variable thereby increasing power; and (2) adjust the dependent-variable means so that they are free of the linear effects attributable to the concomitant variable thereby reducing bias.

Analysis of covariance is often used in three kinds of research situations. One situation involves the use of intact groups with unequal concomitant-variable means and is common in educational research. The procedure statistically equates the intact groups so that their concomitant-variable means are equal. Unfortunately, a researcher can never be sure that the concomitant-variable means that are adjusted represent the only nuisance variable or the most important nuisance variable on which the intact groups differ. Random assignment is the best safeguard against unanticipated nuisance variables.

Analysis of covariance also can be used to adjust the concomitant-variable means when it becomes apparent that although the participants were randomly assigned to the treatment levels, the participants at the beginning of the experiment were not equivalent on a nuisance variable. Finally, analysis of covariance can be used to adjust the concomitant-variable means for differences in a nuisance variable that develop during an experiment.

Statistical control and experimental control are not mutually exclusive approaches to reducing error variance and minimizing the effects of nuisance variables. It may be convenient to control some variables by experimental control and others by statistical control. In general, experimental control involves fewer assumptions than statistical control. However, experimental control requires more information about the participants before beginning an experiment. Once data collection has begun, it is too late to randomly assign participants to treatment levels or to form blocks of matched participants. The advantage of statistical control is that it can be used after data collection has begun. Its disadvantage is that it involves a number of assumptions such as a linear relationship between the dependent variable and the concomitant variable that may prove untenable.

REFERENCES

Anderson, N.H. (2001) *Empirical Direction in Design and Analysis*. Mahwah, NJ: Lawrence Erlbaum Associates.

Dean, A. and Voss, D. (1999) *Design and Analysis of Experiments*. New York: Springer-Verlag.

Finney, D.J. (1945) 'The fractional replication of factorial arrangements', *Annals of Eugenics*, 12: 291–301.

Finney, D.J. (1946) 'Recent developments in the design of field experiments. III. Fractional replication', *Journal of Agricultural Science,* 36: 184–191.

Fisher, R.A. (1925) *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd.

Fisher, R.A. (1935) *The Design of Experiments*. Edinburgh and London: Oliver and Boyd.

Giesbrecht, F.G., and Gumpertz, M.L. (2004) *Planning, Construction, and Statistical Analysis of Comparative Experiments*. Hoboken, NJ: Wiley.

Hacking, I. (1983) *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science*. Cambridge, England: Cambridge University Press.

Kempthorne, O. (1947) 'A simple approach to confounding and fractional replication in factorial experiments', *Biometrika*, 34: 255–272.

Kirk, R.E. (1995) *Experimental Design: Procedures for the Behavioral Sciences* (3rd edn.). Pacific Grove, CA: Brooks/Cole.

Kirk, R.E. (2005) 'Analysis of variance: Classification', in B. Everitt and D. Howell (eds.), *Encyclopedia of Statistics in Behavioral Science*, 1: 66–83. New York: Wiley.

Masling, J. (1966) 'Role-related behavior of the subject and psychologist and its effects upon psychological data', in D. Levine (ed.), *Nebraska Symposium on Motivation. Volume 14.* Lincoln: University of Nebraska Press.

Maxwell, S.E. and Delaney, H.D. (2004) *Designing Experiments and Analyzing Data* (2nd edn.). Mahwah, NJ: Lawrence Erlbaum Associates.

Orne, M.T. (1962) 'On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications'. *American Psychologist*, 17, 358–372.

Rosenberg, M.J. (1965) 'When dissonance fails: on eliminating evaluation apprehension from attitude measurement', *Journal of Personality and Social Psychology*, 1: 18–42.

Rosenthal, R. (1963) 'On the social psychology of the psychological experiment: The experimenter's hypothesis as an unintended determinant of experimental results', *American Scientist*, 51: 268–283.

Rosenthal, R. (1969) 'Interpersonal expectations: Effects of the experimenter's hypothesis', in R. Rosenthal and R.L. Rosnow (eds.), *Artifact in Behavioral Research.* New York: Academic Press. pp. 181–277.

Rosenthal, R. (1978) 'How often are our numbers wrong?' *American Psychologist*, 33: 1005–1008.

Ryan, T.P. (2007) *Modern Experimental Design*. Hoboken, NJ: Wiley.

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

Sheridan, C.L. (1976) *Fundamentals of Experimental Psychology* (2nd edn.). Fort Worth, TX: Holt, Rinehart and Winston.

Yates, F. (1937) 'The design and analysis of factorial experiments', *Imperial Bureau of Soil Science Technical Communication* No. 35, Harpenden, UK.