CONTENTS

6 Randomized Experimental Designs

> *It is the method of reasoning, and not the subject matter, that is distinctive of mathematical thought. A mathematician, if he is of any use, is of use as an expert in the process of reasoning, by which we pass from a theory to its logical consequences, or from an observation to the inferences which must be drawn from it.* Sir Ronald Fisher

At the beginning of the 20th century, Mill's methods defined the concept of control in experimental design. But, as we discussed in Chapter 3, there are problems in applying Mill's methods in experimental work in the social sciences because in these fields it is impossible to achieve the perfect control the methods require. In psychology, there always will be uncontrolled variables introduced by the many differences among people—in attitudes, personality, abilities, and prior experiences.

Given such uncontrollable variables, how can experimenters reach valid conclusions from the results of experiments? We discussed one solution to this problem, involving correlational analysis, in the last chapter; if the uncontrolled variables can be measured, their effects can be corrected for by statistical methods. In this chapter, we consider a second solution to the problem of uncontrolled variables, one developed by Sir Ronald A. Fisher, a British scientist, who was a follower of Galton's ideas in eugenics and a colleague of Pearson and Yule.

Fisher's innovative experimental designs incorporated the controls of Mill's methods and introduced the new technique of randomly assigning subjects to treatments. Today his methods are the standard of excellence for experimental research.

Fisher was a child prodigy in mathematics. He graduated from Cambridge University in England in 1913, with concentrations in mathematics and the new field of genetics. When World War I interrupted his scientific career, Fisher was excluded from military service because his eyesight was poor. So he did "war work" instead, and took the teaching position of another man who went to war. On Armistice Day, he quit teaching, which was not to his liking, and started looking for another job.

Fisher was considering two very different careers—subsistence farming, an occupation that would let him live an "ideal eugenic life" and raise a large family, and research in the new field of biometry, the application of mathematics to biology and genetics. While deliberating on these radically different alternatives, he heard of an available position for a statistician at the experimental agriculture research station at Rothamsted. Although this was just a temporary position, analyzing data already collected at the station, the unusual combination of mathematics and farming must have attracted him, because he decided to set aside his dream of subsistence farming and take the position. Fisher's choice was fortunate for science.

The director of the research station soon realized Fisher's immense talent and the temporary job was made long-term: "It took me a very short time to realize that he was more than a man of great ability, he was in fact a genius who must be retained" (Box, 1978, p. 97). Within a few years, Fisher would develop a remarkable theory of experimentation, complete with experimental designs and a method of data analysis, which he called the analysis of variance.

Fisher's experimental designs were presented in a 1926 paper entitled "The Arrangement of Field Experiments." In this paper Fisher developed the logic and the advantages of his new methods. We will introduce his methods by closely following the examples from agriculture that he used in that paper; these examples illustrate the logic of the methods especially clearly. Once the basic designs are discussed, we will go on to consider how they are applied in psychology.

6.1 A MEASURE OF ERROR

Imagine, as Fisher did, a large field divided into two equal plots. Wheat is planted in both plots and the plots are treated exactly the same except that one is fertilized and the other is not. For convenience, let's refer to the fertilized plot as Al and the other plot as A2. The experimenter wants to discover the effect of the fertilizer on wheat yields.

The design of this experiment follows the logic of the method of difference: Only one antecedent is different for the two plots, Al versus A2 (the independent variable), while other variables are controlled. According to Mill's method, if a difference is found in the wheat yields of the two plots (the dependent variable), it would be due to the fertilizer.

Let's say, as Fisher did, that plot Al produces the greater yield. To make the outcome numerical, say that Al yields 100 bushels and A2 yields 82 bushels. Mill's method would lead to the conclusion that the 18-bushel advantage of Al over A2 is due to the fertilizer. However, in practice, we couldn't be confident in drawing this conclusion—because Mill's method could not be applied perfectly. It is not possible to control for every difference between the plots. Plot Al might have better soil than A2, or better drainage, or less insect or bird damage. As Fisher put it, "What reason is there to think that, even if no [fertilizer] had been applied, the [plot] which actually received it would not still have given the higher yield?" (Fisher, 1926, p. 504).

TABLE 1 FARMER'S RECORDS COMPARING WHEAT YIELDS (IN BUSHELS) OF PLOTS A1 AND A2

Year	*Plot A1*	*Plot 2l*	*A1-A2*
1906	88	80	+8
1907	89	87	+2
1908	84	90	-6
1909	95	100	-5
1910	94	92	+2
1911	85	80	+5
1912	80	79	+1
1913	87	83	+4
1914	79	85	-6
1915	87	90	-3
1916	93	92	+1
1917	98	87	+11
1918	98	90	+8
1919	95	97	-2
1920	94	89	+5
1921	86	80	+6
1922	82	77	+5
1923	82	85	-3
1924	91	87	+4

Here, then, is a perfect illustration of the major problem with the method of difference. The advantage of Al could be due to the fertilizer, or due just to uncontrolled events. How can the researcher decide between these possibilities? Fisher considered two types of evidence that might help. First, what if the farmer stated that he chose the plots fairly and had no reason to believe that one plot had better soil than the other? Or second, what if the farmer had kept records over several years of the wheat yields of these two plots? Fisher dismissed the farmer's opinion as evidence, since it could not be substantiated, but he felt the records would provide valuable information.

Let's say the farmer had records for the past 19 years of the wheat yields for both plots *without fertilizer on either plot*. Table 1 shows these yields; the difference in the yields of the plots also is shown there and plotted as a histogram in Figure 1.

The results for the first 19 years show the differences in the yields of plots Al and A2 when the plots were treated *uniformly;* these differences would be due to uncontrolled variables. The differences vary from +11 bushels (the greatest advantage for Al) to -6 bushels (the greatest advantage for A2). Now compare these differences with the result of the experiment, which is marked on the histogram. Not once in the 19 years did Al have an advantage as large or larger than the 18-bushel advantage that occurred when Al was fertilized.

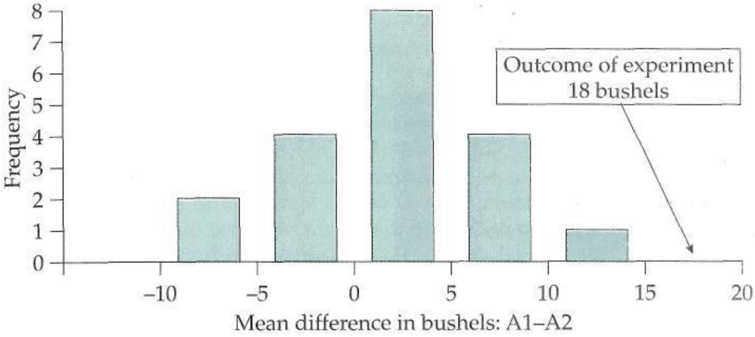


FIGURE 1
Histogram of farmer's records for 19 years prior to the experiment.

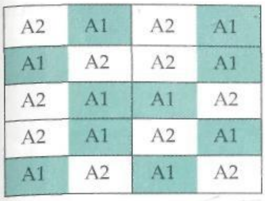
On the basis of this finding, Fisher concluded: "Either there is something in the treatment, or a coincidence has occurred such as does not occur more than once in twenty trials" (Fisher, 1926, p. 504).

What kind of coincidence? The coincidence would be that the year selected for the experiment just happened to be the year of the maximum advantage observed for plot Al, the actual effect of the fertilizer being nil. If the fertilizer is ineffective, the probability of this coincidence is 1/20, or 5%. Since Fisher judged this probability small enough to argue against the hypothesis that the fertilizer was ineffective, he concluded that the fertilizer was effective.

With this design, Fisher offered a solution to the problem of the method of difference. By collecting additional information on the results that would occur due just to uncontrolled variables, the researcher would be able to make a reasoned decision about the effectiveness of the treatment. Using this design, however, it would take many years to complete even a simple experiment. But Fisher didn't intend this method as a practical model for research. In presenting this design, he wanted only to show the kinds of information needed to interpret the outcome of the experiment. He argued that:

> what is required to interpret the outcome of an experiment is a *valid measure of error,* a measure of which outcome will occur due to uncontrolled variables.

This is the heart of Fisher's approach—finding a valid measure of error. Mill had argued for perfect control in the experiment, so that errors could be reduced to zero. Because Mill's solution is impossible to achieve in practice, Fisher proposed to *measure the extent of the error* instead, and to use this measure in interpreting the outcome of the study. From Fisher's viewpoint, if you can't eliminate all error, the best alternative is to measure it, so that you can take account of the error in drawing your conclusions.



A2	A1	A2	A1
A1	A2	A2	A1
A2	A1	A1	A2
A2	A1	A2	A1
A1	A2	A1	A2

FIGURE 2
Randomized block design with two treatments, A1 and A2, and 10 blocks.

Fisher went on to show how to derive this measure of error from experiments that could be done in a single growing season. We will discuss his designs with one independent variable in this chapter and devote Chapter 8, Factorial Designs and Interactions, to the more complex designs with more than one independent variable.

6.2 THE RANDOMIZED BLOCKS DESIGN

Fisher's *randomized blocks design,* now a standard in psychological research, was used first in agriculture. The researcher would divide the field chosen for the experiment into a number of smaller areas, called *blocks*. Let's pick 10 as a number to work with. Each of the 10 blocks would be further divided into two plots. Then, within each block one plot would be *randomly selected,* say, by a coin toss, to receive treatment Al, the fertilizer, and the other would receive no fertilizer, treatment A2. Figure 2 shows a randomly selected arrangement of the treatments in the field. Notice that treatments Al and A2 appear in each of the 10 blocks, but their positions within the block vary randomly.

6.2.1 Replication

At the end of the growing season, the crops on each of the 20 plots would be harvested and their yields measured and recorded. As in the previous design, half of the total area of the field would be treated with fertilizer and half would not. But now, instead of comparing the yields in the two halves of the field, we can compare yields within each of the 10 blocks. In effect, the original experiment is *replicated* 10 times using smaller plots, and the method of difference is applied 10 times, once in each block. This replication, an innovation of Fisher's, is necessary to measure the error in the experiment.

> *Replication,* a major feature of all of Fisher's designs, was not present in any of Mill's methods.

6.2.2 Random Assignment

The *random assignment* of treatments to plots (or randomization) is the second major feature of this design.

> With random assignment, each plot has an equal probability of receiving each treatment.

randomization is done for two reasons. First,

> randomization avoids any bias that may occur if a nonrandom or systematic assignment is used.

If, for example, representatives from the fertilizer company made the assignment, they might select the better looking soil to receive the fertilizer, thereby creating a bias in the study. Even a neutral observer might bias the assignment unconsciously. Also, a systematic assignment, such as alternating the treatments in successive plots (for example, Al A2 Al A2), could bias the study if uncontrolled soil conditions in the field also had this pattern of variation.

The second reason for using random assignment is that

> random assignment is necessary to determine a valid measure of error for the experiment, a measure of which outcomes to expect due to uncontrolled variables.

Without this measure of error, there is no good way to interpret the results of the study.

> Random assignment, like replication, was not used in Mill's methods.

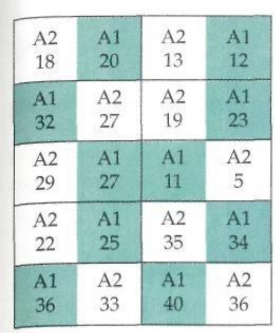
Fisher felt that randomization was so critical for experimentation that he and a colleague published tables of random numbers to make it easy for researchers to randomly assign treatments (Fisher & Yates, 1953).

6.2.3 Determining the Measure of Error

Now let's look closely at the outcome of the experiment to see why randomization and replication are necessary to calculate a measure of error. Figure 3 shows the wheat yields in bushels for each of the 20 plots in the field.

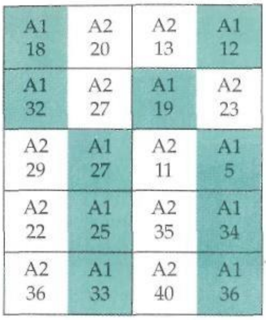
The average yield for the plots getting Al is 26.0 bushels; the average yield for A2 is 23.7 bushels. These means were calculated by adding the yields for each type of plot and dividing by the total number of plots getting that treatment, 10 in this case. Since the distribution of yields across different plots is expected to have a normal distribution, the mean is the appropriate summary statistic.

The results show that the mean difference between the fertilized and unfertilized plots is 26.0 - 23.7 = 2.3 bushels per plot, the advantage going to the fertilized plots. The question is whether this difference is due to the fertilizer or uncontrolled variables, such as soil fertility or bird damage? To answer this question, we need a measure of error. Fisher's solution to this problem was ingenious.



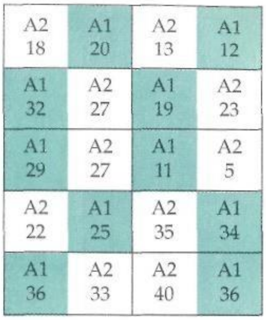
A2 18	A1 20	A2 13	A1 12
A1 32	A2 27	A2 19	A1 23
A2 29	A1 27	A1 11	A2 5
A2 22	A1 25	A2 35	A1 34
A1 36	A2 33	A1 40	A2 36

FIGURE 3
Yields in bushels for the 20 plots: Mean A1 – mean A2 = 2.3 bushels.



A1 18	A2 20	A2 13	A1 12
A1 32	A2 27	A1 19	A2 23
A2 29	A1 27	A2 11	A1 5
A2 22	A1 25	A2 35	A1 34
A2 36	A1 33	A2 40	A1 36

FIGURE 4
Relabeling the plots: Mean A1 – mean A2 = –1.5 bushels.



A2 18	A1 20	A2 13	A1 12
A1 32	A2 27	A1 19	A2 23
A1 29	A2 27	A1 11	A2 5
A2 22	A1 25	A2 35	A1 34
A1 36	A2 33	A2 40	A1 36

FIGURE 5
Relabeling the plots: Mean A1 – mean A2 = +1.1 bushels.

Assume, for now, that the fertilizer is completely ineffective. If this is true, what we have done in the study is simply to label plots of the field randomly as Al or A2 and compare the means of plots that have been given these arbitrary labels. If the fertilizer is ineffective, plots Al and A2 actually were treated uniformly, and the observed mean difference of 2.3 bushels would be a result of uncontrolled variables.

With a different assignment of labels, the outcome of the study would have been different. Figure 4 shows a different labeling done following the same scheme of randomization used in the actual experiment.

With this labeling, and assuming the fertilizer does not work, the mean difference between the plots would have been -1.5 bushels. Figure 5 shows yet another labeling that could have happened. With this labeling, the mean difference would be +1.1 bushels.

If we continued relabeling the plots and computing the resulting mean difference, we would end up with a set of values for the mean difference, values that would be expected if the fertilizer did not work. These values show us the mean differences to expect due just to uncontrolled variables.

This set of values is exactly the measure of error we are looking for!

Fisher thought that 500 values would be sufficient to accurately measure the error of the study We did the relabeling 500 times and recorded the resulting mean differences. They are collected and presented in the histogram in Figure 6.

6.2.4 The Null Hypothesis

The histogram shows the mean differences between plots Al and A2 that would be expected if the fertilizer is ineffective.

Fisher called the hypothesis that the fertilizer is ineffective the *null hypothesis*.

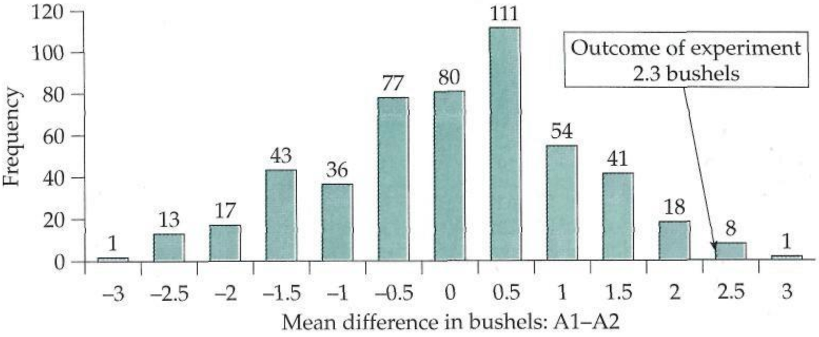


FIGURE 6
Mean differences in yields expected due to uncontrolled variables, A1 – A2: Results of 500 random assignments.

The *null hypothesis* in an experiment states that the independent variable has no effect on the dependent variable.

The histogram shows that the mean differences, given that the null hypothesis is true, range from a low of -3 bushels to a maximum of +3 bushels.

6.2.5 The Significance Probability, p

Now we can compare the outcome of the study with this measure of error. The actual outcome, a mean difference of 2.3 bushels, is marked on the histogram. In the 500 relabelings, only 9 assignments gave a mean difference of 2.3 or more. If the fertilizer is ineffective, the probability of getting a mean difference of 2.3 or more is p = 9/500 = .018.

The statistic p is the *significance probability* for the test of the null hypothesis.

> The *significance probability*, p, is the probability, if the null hypothesis is true, of getting the observed mean difference or an even larger value.

6.2.6 Interpreting p

Fisher contended that the conclusion drawn from a study should depend upon the value of p. He suggested that the cutoff point of p = .05 be used in reaching a conclusion according to the following rule:

If p is less than or equal to .05, $p < .05$, then reject the null hypothesis, since the results are inconsistent with this hypothesis.

If p is greater than .05, p > .05, then do not reject the null hypothesis, since in this case the results are consistent with this hypothesis.

If you remember from Chapter 4, the cutoff point for p is called the alpha (a) level of the test. In our example, p = .018, which is less *than* the a = .05 (5%) cutoff point, so the null hypothesis is *rejected*. The conclusion is that there is good evidence that the fertilizer works.

When the null hypothesis is rejected, the result, the observed mean difference, is said to be *statistically significant* at the 5% *alpha level* or the 5% *level of significance*.

> The *level of significance* is the alpha level, the cutoff point for p in reaching a conclusion about the null hypothesis; 5% is the accepted standard today.

It is important to realize that this evidence, a statistically significant mean difference, is *not proof* that the fertilizer worked, just *good evidence* that it worked.

> A statistically significant result is one that is unlikely to occur due just to the uncontrolled variables in a study.

But it may have occurred. In fact, if the fertilizer is ineffective, the conclusion of the statistical test will be wrong 5% of the time. Because of the possibility of this error in interpretation, it is necessary to replicate experimental results in many different studies before they become accepted as fact. According to Fisher:

> A scientific fact should be regarded as experimentally established only if a properly designed experiment *rarely fails* to give this [5%] level of significance. (Fisher, 1926, p. 504)

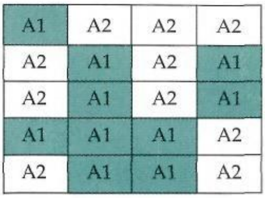
The randomized blocks design is a practical design for agriculture that takes only a single growing season to yield results. The statistical procedure of doing 500 relabelings, called a *randomization test* or a *Monte Carlo test,* is practical today because we have computers. In the 1920s, it took hours of boring work by hand.

6.2.7 The t Test

Fisher recommended using an approximation to the randomization test to avoid these laborious computations. In 1908, William Gosset, a colleague of Pearson's, publishing under the name "Student," developed a statistical test, called the *t test* (Student, 1908).

> The *t test* is used to test for differences between the means of two groups in a study in which the subjects are randomly sampled from a large (actually infinite) population.

Even though in the randomized blocks design the subjects (plots of land) are not randomly selected from a large population, Fisher showed that Student's *t* test was a good approximation to the randomization test.



A1	A2	A2	A2
A2	A1	A2	A1
A2	A1	A2	A1
A1	A1	A1	A2
A2	A1	A1	A2

FIGURE 7
Completely randomized design with two treatments, Al and A2, and with 20 plots.

Student's *t test* is quick and easy to compute. A value of the statistic, called i, is computed from the results of the study, and this value can be looked up in a table to see if the mean difference is significant. The computations and use of the table are presented in statistics texts. Today, even though modern computers make calculating Fisher's preferred test, the randomization test, easy, the t test is still the more popular procedure. In most cases, the choice between these two tests is, as they say, "academic," since their results agree closely; in the example computed above, the randomization test gives $p = .018$; the t test gives p = .013.

If you go to any shopping mall today, you will be able to find inexpensive pocket calculators that have the t test built in; the user only has to enter the data from a study and the computer does the entire computation. Student would have been shocked to learn how popular his test would become. Its popularity is due to the fact that it can be used with Fisher's designs.

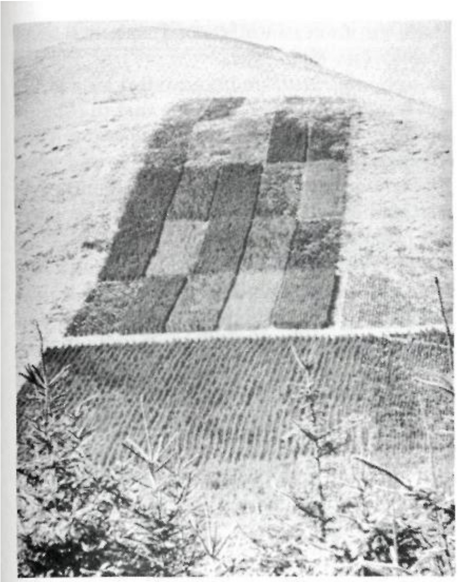
6.3 COMPLETELY RANDOMIZED DESIGN AND THE LATIN SQUARE DESIGN

In the *completely randomized design,* instead of using the blocking of the randomized blocks design, the treatments are randomly assigned to plots throughout the whole field. The field is divided into a number of plots and the treatments are randomly assigned to them. When two treatments, Al and A2, are being compared, half the plots are randomly assigned to Al and the other half go to A2. Figure 7 shows a completely randomized design with 20 plots.

Because the pattern of randomization is different in this design than in the randomized blocks design, the computations of the statistical test also are different. Fisher developed a modification of Student's t test, called the t test for independent groups, as an approximation to the randomization test for this case.

As we will see, the completely randomized design is popular in psychology, but it is used infrequently in agriculture because of the possibility of an apparent imbalance in the assignment of treatments to plots. If most of the

Al treatments were assigned by chance to the north section of a field and most of the A2 treatments ended up in the south section, there would be an apparent bias in the study. No experimenter would want to conduct an agricultural study with this assignment. The randomized blocks design avoids this type of apparent bias by restricting the random assignments to within blocks, guaranteeing that both treatments are applied evenly all over the field.



BETTGELERT FOREST LAYOUT

Elevation:					
1730–1800′	B	A	E	D	C
1530–1730′	C	E	B	A	D
1460--1590′	A	C	D	E	B
1340–1460′	D	B	A	C	E
1250–1340′	E	D	C	B	A

A. Sitka spruce
B. Japanese larch
C. Sitka spruce/Japanese larch 50/50
D. Sitka spruce/Pinus contorta 50/50
E. Norway spruce/European larch 50/50

Two rows of Beech planted on each side of the series.

FIGURE 8
Photo taken at Bettgelert Forest 16 years after the trees were planted, with the layout of the planting.

Fisher's *Latin square design* also uses very restricted randomization to avoid apparent bias in assignments. In the Latin square design, the field first is divided into rows and columns, the number of rows and columns corresponding to the number of treatments being compared. With 2 treatments, there would be 2 rows and 2 columns; with 5 treatments, there would be 5 rows and 5 columns. Treatments are assigned to this grid so that one and only one treatment falls in any given row or column.

Figure 8 illustrates how the Forestry Commission in Wales used a 5 x 5 Latin square to study the effects of altitude on the growth of different varieties of pine trees at Bettgelert Forest in 1929. The researchers laid out the Latin square on a steep hill so that each row was at a different elevation. Soil fertility at each altitude was controlled by randomly assigning different varieties and combinations of trees to different positions in the row. The photograph, taken 16 years after planting, clearly shows the differential effects of altitude on the trees.

Fisher published tables giving all possible arrangements of treatments for Latin squares of sizes 2, 3, 4, 5, and 6. The experimenter would lay out the field in the appropriate number of rows and columns and then pick at random an arrangement of treatments from Fisher's table. This randomization ensures that there will be no systematic bias and provides the basis for the measure of error in the study. Again, since the pattern of randomization is different from the previous designs, the calculations of the statistical test also are different.

The computations of the Latin square are found in advanced statistics texts and in handbooks of experimental design (see Kirk, 1982).

The Latin square is a popular design in agriculture because it does a good job of controlling for gradients of soil fertility in a field. Imagine that the soil in the field is best at the north end and becomes progressively worse going from north to south. In the Latin square design each treatment appears in each row and column, so each treatment would be equally applied to the good and poor soil, thus controlling for soil quality. Gradients of soil fertility are common, so the Latin square design is well suited for agricultural research.

Fisher's randomized designs were revolutionary. When Fisher proposed them, other scientists were recommending systematic designs for agricultural experiments. They were skeptical at first about Fisher's randomization method. Student, for example, thought that some systematic designs would result in a smaller error than randomized designs and therefore would be more sensitive. As late as the 1940s, agriculture texts posed the question of which design, systematic or random, was better (Leonard & Clark, 1939).

Today, systematic designs are not even classified as "true" experiments (see Chapter 10, Field Research). When either type of design is possible, randomized designs are preferred. Randomized designs replaced systematic designs for the reasons Fisher presented in his 1926 paper:

Random assignment avoids any systematic bias in assigning treatments and is essential for calculating a valid measure of error.

6.4 Fisher's Designs In Psychology

Fisher developed his designs for agricultural research. In psychology we are not interested in treating plots of land, and our subjects do not come laid out in a field so that they can be easily blocked, and we do not face problems of bird damage or differences in soil fertility. But there are enough parallels between agricultural studies and psychological experiments to make Fisher's designs the methods of choice for experiments in psychology. Table 2 outlines these similarities.

The subjects in psychological research are people or animals. The independent variable is the type of treatment they receive. The treatments, in psychology, vary widely, from schedules of reinforcement, to differently shaped visual stimuli, to different types of psychotherapy. The dependent variable is a measure of the subjects' behavior in the study, behavior that is thought to be influenced by the independent variable. The dependent variable might be the rate of bar pressing in an animal learning experiment, the perceived intensity of a stimulus in a perception study, or the severity of depression in research evaluating psychotherapy.

In the completely randomized design, subjects are randomly assigned to treatments with the sole restriction being that equal numbers of subjects be assigned to each treatment. (Equal numbers are not necessary but do lead to the most sensitive design; see the discussion of power later in this chapter.)

TABLE 2 COMPARISON OF AGRICULTURAL AND PSYCHOLOGICAL EXPERIMENTS

Term	*Agriculture*	*Psychology*
Subjects	Plots in field.	People (or animals) who meet the criteria for inclusion.

Independent variable	Different treatments applied to the soil	Different treatments given to the subjects.
Dependent variable	Measure of yield of crops, e.g., wheat yield in bushels	Measure of subjects' behavior, e.g., severity of depression
Random assignment	Random assignment of plots to treatments.	Random assignment of subjects to treatments.
Completely randomized design	Plots in field are randomly assigned to treatments with the only restriction being that equal numbers of plots are assigned to each treatment.	Subjects are randomly assigned to treatments with the only restriction being that equal numbers of subjects are assigned to each treatment.
Randomized Blocks design	The field is divided into blocks, blocks are subdivided into plots, and then plots within a block are randomly assigned to treatments. Random assignment is restricted to within blocks.	Subjects are divided into blocks. A block is a group of subjects who are similar to each other on specified criteria. Subjects within each block are randomly assigned to treatments. Random assignment is restricted to within blocks.
Latin square design	Randomization is restricted to preset patterns of applying treatments to the plots. The patterns have each treatment in the study in each "row" and "column" of the field.	Randomization is restricted to preset patterns of assigning subjects to treatments. The design is used in psychology to test for effects within subjects.
Conditions held constant	Application of seed, preparation of soil, duration of study, method of harvest, amount of watering and weeding, etc.	Initial description of study to subjects, duration of study, instructions, methods of

		measuring dependent variables, etc
Uncontrolled events	Bird damage, weather, insect damage, etc.	Equipment failure, experimenter mistakes in the protocol, fire alarm during study, missed appointments, etc
Uncontrolled differences among subjects	Differences in soil fertility, water drainage, etc	Differences in personality, abilities, interests, past history, etc.

One way to assign subjects randomly to treatments is to write their names on pieces of paper, put the pieces in a hat, shake well, pick out half the subjects for one treatment, and assign the remaining subjects to the other treatment.

Another method of randomly assigning subjects to treatments is by computer. We have included computer programs for random assignment in Chapter 12, Planning the Study.

The randomized blocks design uses more restricted random assignment to groups than the completely randomized design. In agriculture, adjacent plots of land form a block and the randomization takes place within blocks. Because adjacent plots should be more similar in soil fertility than nonadjacent plots, blocking helps to control for differences in fertility. The best parallel to adjacent plots in psychology would be identical twins. Each pair of twins would be one block; the study would have several blocks. Within each block, one twin would be randomly assigned to one treatment and the other would get the second treatment. Since twins are similar in many ways, the study would achieve good control over differences among the subjects.

Twin studies are rare because so few twins are available. But, as the following experiment illustrates, when twin studies can be done they often are models of control. One recent well designed experiment in medicine, for example, may change the way parents feed their children (Johnston et al., 1992). This 3-year study examined the effects of calcium supplements on the bone density of 70 pairs of identical twins, ages 6 to 14. One twin in each pair was randomly chosen to receive 1,000 mg of calcium daily; the other received a placebo that only looked and tasted like the calcium supplement. The results showed that the supplements increased the children's bone density.

If twins are not available, the next best alternative is for researchers to form the blocks themselves.

> Subjects can be paired, or matched, based on their similarity on variables that the experimenter wishes to control.

Matching might be done on sex, age, education, ability, or degree of illness, for example. Random assignments then would be made by selecting subjects for each of the treatments from within these matched blocks of subjects.

In agricultural research using the Latin square design, the treatments are assigned to different plots of a field so that each treatment falls only once in each "row" and "column" of the field. In psychology, there is no single parallel to the rows and columns of a field.

> A common application of the Latin square in psychology is in experiments where each subject receives all of the treatments at different times.

Let's say a psychologist wants to study the effects of caffeine on cognitive functioning, as measured by performance on simple arithmetic problems. The researcher decides to use a within-subjects design. Each subject receives each of four doses of caffeine: no caffeine, dose 0; low caffeine, dose 1; medium caffeine, dose 2; and high caffeine, dose 3. To minimize carryover effects, a one-day interval is planned between the doses.

Figure 9 shows a Latin square design for this study. The columns of the square correspond to the orders, 1st, 2nd, 3rd, or 4th, of administering the doses. The rows correspond to groups of subjects. All the subjects in a group receive the doses of caffeine in the same order. The subjects are randomly assigned to the groups. The entries in the square show the specific doses. The subjects in Group 1, for example, are tested first on dose 2, then dose 0, followed by dose 3, and finally, on the fourth day, dose 1. This particular arrangement of doses was picked at random from a table of possible 4x4 Latin squares (Fisher & Yates, 1953).

Order

	1st	2nd	3rd	4th
Group 1	dose 2	dose 0	dose 3	dose 1
Group 2	dose 0	dose 1	dose 2	dose 3
Group 3	dose 1	dose 3	dose 0	dose 2
Group 4	dose 3	dose 2	dose 1	dose 0

Figure 9. Latin square design showing the order of giving 4 different doses of caffeine to 4 groups of subjects.

The Latin square balances the order of administering the caffeine. Each dose is given to one group of subjects in each possible order; that is, dose 0 is presented first to one group; dose 1 is presented first to another group, etc. This helps to control for carryover effects and to balance out any effects of practice on test taking.

In the next section, we present three psychology experiments to illustrate the completely randomized design, the randomized blocks design, and the Latin square. The Rush, Beck, Kovacs, and Hollon experiment (1977) was the first to find that a psychotherapy was better than a standard drug therapy for a major psychiatric disorder. The study by Held and Hein (1963) established a basic fact about the development of visual perception. The findings of the study by Wolraich et al. (1994) contradict conventional wisdom about the effects of sugar on children's behavior.

6.4.1 Completely Randomized Design: Evaluating Cognitive Therapy

In the mid 1970s, cognitive psychotherapy was a promising treatment for depression. Clinical experience treating patients was positive; however, the effectiveness of the therapy had not been tested in a controlled experiment using random assignment of patients to treatments. In 1977, Rush et al. published the first such test.

The subjects were depressed patients referred to the University of Pennsylvania Hospital for treatment. To participate in the study, patients had to meet rigorous inclusion criteria, including moderate to severe levels of depression, a diagnosis of depressive syndrome based on published criteria, no history of schizophrenia or alcoholism, and no contraindications for antidepressant medication. Over 110 applicants were screened to find the 41 patients included in the study.

The patients signed a consent form agreeing to receive either cognitive therapy or drug therapy with imipramine, a tricyclic antidepressant (a standard drug therapy for depression). Then they were randomly assigned to one of the therapies. The severity of each patient's depression was monitored throughout the 12-week treatment using three measures: the Beck Depression Inventory, the Hamilton Rating Scale for Depression, and the Raskin Depression Scale.

At the end of the treatment, the mean depression scores on all three measures of patients in the cognitive therapy group were significantly lower than those in the drug therapy ($p < .05$). This study, published in 1977, in volume 1 of a new journal, *Cognitive Therapy and Research,* inspired many other studies of cognitive therapy.

6.4.2 Randomized Blocks Design: Stimuli Necessary for Perceptual Development

By the early 1960s, there was evidence that normal vision in cats depends upon their experiencing varied visual stimulation as kittens. Kittens deprived of normal stimulation, either by being physically restrained or by having their eyes covered with hoods that let in only diffuse light, later showed visual deficiencies when compared to litter mates raised normally. Based on these results, Richard Held and Alan Hein (1963) considered two alternative hypotheses about the kind of stimulation needed for normal visual development. According to one hypothesis, stimulation received while the animal is passive would be sufficient to produce normal vision. According to the second hypothesis, young animals must be free to *create changes* in their visual stimulation through their *own movements* for normal vision to occur.

To decide between these hypotheses, Held and Hein tested 8 pairs of kittens, each pair from a different litter. All 16 kittens were raised in darkness until they were strong enough to be in the study (at 8 to 12 weeks). Then they were exposed to carefully controlled visual stimulation for three hours a day. One kitten in each pair was randomly assigned to the "Active" condition and the other to the "Passive" condition. The active member (A) of the pair was allowed to walk inside an illuminated circular pen that was 4 feet in diameter with 1-inch-wide black-and white vertical stripes on its wall. The passive member (P) was placed on the other side of the pen from A, in a physical apparatus with rods, gears, and pulleys that operated so that when A moved, P moved an equivalent distance.

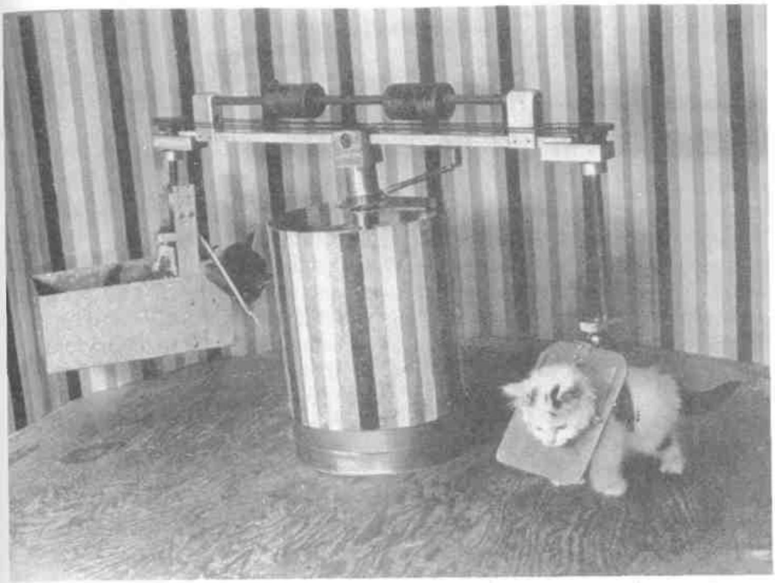
The apparatus permitted P to make only slight head and eye movements on its own. By this means the visual stimulation of both kittens was kept nearly equal; but A's stimulation was self-produced, whereas P's was not. Each pair thus provided a test between the two hypotheses.

This study involved two types of matching. First, the kittens in each pair were litter mates; so they were expected to be more similar to each other than unrelated kittens. Second, the kittens in each pair were "yoked" together; that is, they were placed in an apparatus that operated so that the movements of the active kitten controlled the visual stimulation of the passive one.

> Designs in which one subject's behavior controls the outcome of another are called *yoked control designs;* these designs are used to control for variables that are directly affected by the behavior of the subjects themselves.

Held and Hein used this design to control the variety of visual stimulation presented to both kittens, while simultaneously allowing one kitten to be active and one to be passive.

The daily experimental sessions continued until one member of the pair could pass the "paw placement test." In this test, the kitten was carried forward and downward toward the edge of a table; it passed if it showed visually mediated anticipation of contact by extending its paws as it approached the table. As soon as one of the kittens passed, both kittens were given additional visual tests.



Held and Hein's apparatus for controlling visual stimulation of active and passive kittens.

The results confirmed the researchers' expectations: In each pair, the active kitten passed the paw placement test first ($p < .05$) and also performed better on the other two measures of visual development. The authors concluded that the results "provide convincing evidence for a developmental process, in at least one higher mammal, which requires for its operation stimulus variation concurrent with and systematically dependent upon self-produced movement" (Held & Hein, 1963, p. 876). These results provided an impetus to the development of "feedback" toys for human infants, toys that would give babies varied stimulation dependent on their own movement.

6.4.3 Latin Square: High Sugar Diet for Children

Many parents and schoolteachers are convinced that children are overly sensitive to sugar—that sugar creates a "sugar high" that leads to hyperactivity and poor conduct. This idea was tested by Mark Wolraich and his colleagues (1994) in an elaborate study that controlled the total diets of 48 families for a 9-week period.

Two groups of children were recruited by advertisements and by contacting preschool programs: 25 children, 3 to 5 years old, and 23 children, 6 to 10 years old, all of whom were identified by their parents as being sugar sensitive. At the beginning of the study, the researchers removed all food from the subjects' homes and replaced it with food prepared for the study.

Order

	1st	2nd	3rd
Group 1	diet 1	diet 3	diet 2
Group 2	diet 2	diet 1	diet 3
Group 3	diet 3	diet 2	diet 1

Figure 10. Latin square design showing the order of three diets.

Each subject and the subject's family followed three different diets, each for a 3week period: Diet 1 was high in sugar, with no artificial sweeteners; diet 2 was low in sugar, with aspartame (the ingredient in NutraSweet) as a sweetener; diet 3 also was low in sugar, but with saccharin as the sweetener. The order of presenting the diets was balanced, using a 3 x 3 Latin square (see Figure 10).

Clearly, suggestion is a major threat to the internal validity of this study; if the families knew which diet they were on, their child's behavior might be influenced by their strong expectations that sugar causes behavior problems. To guard against this, the subjects, family members, and experimenters testing the children were not told which diet the subjects were on at any time. Also, although the actual diet changed only every three weeks, the appearance of the diet was changed on a weekly basis. Only one parent correctly guessed the order of the diets.

The children were tested weekly on a battery of tests assessing their academic skills, motor skills, and general activity levels. Their parents, teachers, and the experimenters also rated them on behaviors such as conduct, hyperactivity, and aggression.

The data were analyzed by averaging the scores on these measures during the periods of the three diets. The results showed virtually no differences in the children's behavior associated with diet. The experimenters concluded that "neither sucrose nor aspartame produces discernible cognitive or behavioral effects in normal preschool or in school-age children believed to be sensitive to sugar" (Wolraich et al., 1994, p. 306). We are left with the mystery of why so many parents are convinced that sugar is a factor in the misbehavior of their children.

6.5 POWER ANALYSIS: DECIDING ON THE NUMBER OF SUBJECTS

6.5.1 Type I and Type II Errors

On the basis of their results, Rush et al. concluded that cognitive therapy was more effective than drug therapy for the patients in their study. But this conclusion could be wrong. Rush et al. did their statistical test at the 5% level of significance. This means that there is a probability of 5% that they could conclude that one treatment is more effective than the other when, in fact, the treatments do not differ in effectiveness (that is, when the null hypothesis is true). This error is called a *Type I error*.

> A *Type I error* occurs when the null hypothesis is rejected when it is true.

Using the 5% level of significance, 5 out of 100 experiments will reach an incorrect conclusion that there is an effect of the experimental treatment when there actually is no difference in the treatments.

As a test of this possibility, Rush's study was replicated by Irene Elkin and her colleagues (1989), who found no significant difference between cognitive therapy and drug therapy. But this finding also may be in error. Elkin's group may have committed the error of *not rejecting* the null hypothesis when it is false, a *Type II error*.

> *A Type II error* occurs when the null hypothesis is not rejected when it is false.

Experimenters can never know whether they have made a Type I or Type II error since they can never know the *actual* or *true* effects of the treatments. Experimenters know only the observed results, which always are subject to error. As we have discussed, the probability of making a Type I error is controlled by the statistical test. Doing the test at the 5% level of significance means that the probability of a Type I error is exactly 5%. The Type II error is controlled by the design of the study. A well-designed study with good measures and enough subjects will have a low probability of making a Type II error. This means that the probability will be high of detecting a difference between the treatments if such a difference exists.

The probability of drawing this correct conclusion is called the *power* of the statistical test.

> The *power* of a statistical test is the probability of correctly rejecting the null hypothesis. The power is equal to one minus the probability of a Type II error.

The concept of power was introduced by Jersey Neyman and Egon Pearson; both men were colleagues of Fisher in the late 1920s. Egon Pearson was the son of Karl Pearson, Galton's colleague.

A power of .50 or 50% (we will express power as a percent to avoid the decimal) means that the probability is 50-50 (the same as getting heads on the flip of a coin) that the experiment will detect a true difference between the treatments. A power of 90% means that the probability is 90% that an actual effect will be detected. Other things being equal, the experimenter wants the power to be as high as possible.

Setting the number of subjects in the study is the primary method the experimenter has of controlling the power of an experiment: The more subjects, the greater the power. The power can be set as close to 100% as the experimenter wants by including enough subjects. But large-sized experiments can be expensive to conduct and time-consuming to administer. In addition, it may be difficult or impossible to find enough subjects for a large study. So experimenters must strike a balance between these practical concerns and power in deciding on the number of subjects.

Successful studies with a completely randomized design are possible with a wide range of numbers of subjects—from just a few subjects in a group, to as many as 11,000 subjects per group, for example, in an experiment on the relationship between taking aspirin and heart attacks (Steering Committee of the Physicians Health Study Research Group, 1988).

Deciding on the number of subjects is critical for experimenters. Too few subjects and the study can be a complete washout with no significant findings. When this happens an effective treatment may be overlooked. Too many subjects wastes time and money that could be better spent investigating other aspects of the problem. To make a reasonable decision on the number of subjects, experimenters must understand the relationship between the design of an experiment and power.

6.5.2 The Treatment Effect Size

The power of a statistical test depends not just on the number of subjects, but also on the size of the treatment effect. If the effect is strong (e.g., if everyone getting the treatment changes dramatically and no one in the control group changes), an experiment with only a few subjects will lead to the correct conclusion that the treatment is effective. If the treatment effect is slight, however, many more subjects will be needed to ensure a good likelihood of reaching the correct conclusion.

The fact that power depends on the actual size of the treatment effect presents a problem in designing studies. The experimenter planning the study does not know the size of the treatment effect, since, of course, the study is designed to find this out. So the experimenter must estimate the effect size and plan the study accordingly. Let's look at an example and see how this is done.

Consider an experiment to study the effectiveness of a training program for increasing scores on the SAT verbal exam. The subjects, high school seniors, are randomly assigned to *two* groups. The treatment group receives the training; the control group gets no training. After the training is completed, the subjects in both groups take the SAT exam. In the data analysis, the experimenters plan to compare the mean SAT scores for the two groups using the *t* test at the 5% level of significance. They know that in a national sample the mean SAT

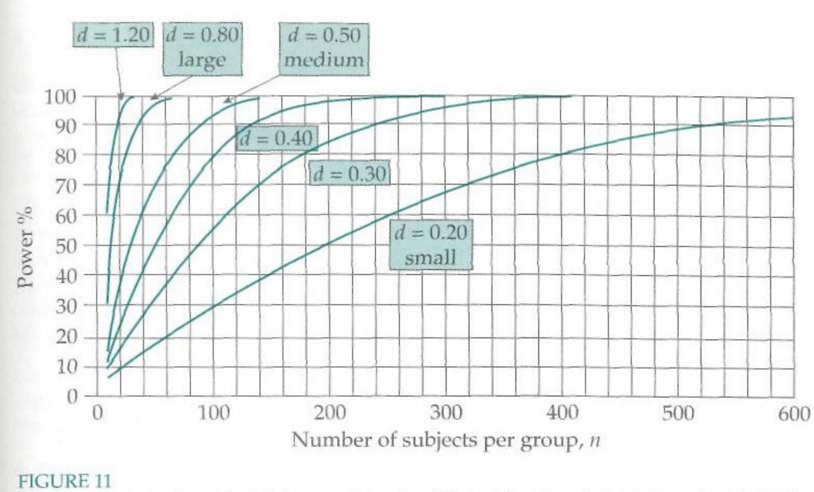


FIGURE 11
Power versus number of subjects per group for different treatment effect sizes, d = MD/SD.

verbal score is set to 500, with a standard deviation of 100. Now, how many subjects should they observe in each group? 10? 20? 50? 100?

Let's consider three possibilities for the size of the training effect.

1. A small sized effect: Let's say SAT scores in the treatment group average only 20 points higher than scores in the control group. Since the standard deviation of the SAT is 100 points, this would be an increase of only 20/100 = 0.20 of a standard deviation. An average student, at the 50th percentile without training, would be at the 58th percentile with training. (The 58th percentile is 0.20 standard deviations above the mean on a normal curve.)

Dividing the mean difference by the standard deviation, as we just did to get 20/100 = 0.20, gives a good measure of the size of a treatment effect. This statistic is commonly used in the analysis of power and is called the *standardized mean difference (or Cohen's measure of effect size), d.*

2. *A* medium-sized effect: Scores are raised by 50 points; d = 50/100 = 0.50. The average student before training would be at the 69th percentile after training.
3. A large-sized effect: Scores are raised by 80 points; d = 80/100 = 0.80.
   The average student before training would be at the 79th percentile after training.

Let's first consider the small effect (d = 0.20). The graph in Figure 11 shows the relationship between power (in %) and the number of subjects per group, n, for different sized treatment effects.

Find the curve for d = 0.20. Next, find the point on the curve where it intersects the 50% power value and read that the corresponding number of subjects is about 190. This means that 190 subjects per group, a total of 380 subjects for the study, are needed to have a 50% chance of detecting this small effect. It probably would not be worthwhile to do the study with a power this low, since half the time you would not expect significant results.

Jacob Cohen (1988), who has been advocating power analysis to psychologists planning research, suggests doing studies with a power of at least 80%. With a small effect size of d = 0.20, this would take about 400 subjects per group. Unless you have access to lots of subjects, doing a study to detect a small effect might be a waste of effort.

With a medium-sized effect (d = 0.50), many fewer subjects would be needed. The graph shows that for a power of 80%, it would take about 65 subjects per group, and for a power of 90%, about 85 subjects per group. For a strong effect, d = 0.80, 25 subjects per group would yield a power of 80%; 35 subjects would give about 90% power.

So what is the bottom line on how many subjects to use? With a medium-sized effect, plan on around 65-85 subjects per group to have a high-powered study. If the size of the effect is large, the study could use as few as 25 subjects per group and still have high power. On the other hand, if you expect a small treatment effect, plan on a large number of subjects (800+). If this is not feasible, consider going back to the drawing board and improving the treatment, or try one of the strategies discussed below to increase power.

6.5.3 How to Estimate the Effect Size in Your Own Research

Estimating the effect size is straightforward if you are doing research on a problem that has been studied before. For example, there is extensive literature on the effectiveness of different treatments for depression, from psychotherapies to electro-convulsive therapies. If you were interested in doing research on depression, you could get a good idea of what effect sizes to expect in your own research by studying the available literature. Virtually all studies publish means and standard deviations on the outcome measures for each treatment group, so calculating the effect size (d = MD/SD) is simple.

If you can find no previous research on your problem, then the power analysis is difficult because you have no basis for estimating the effect size. In such cases, one approach would be to do a *pilot study:*

> *A pilot study* is a small-scale rehearsal of the actual study to test procedures and practice interacting with the subjects.

The results from a pilot study will give you a rough idea of the standard deviation of your measure and the differences to expect between groups. Without a pilot study, your best bet would be to make an educated guess about the effect size, or simply plan for a medium-sized effect and include

40 to 50 subjects per group. This would give you high power for a strong effect, good power for a medium effect, and poor power for a weak effect.

6.5.4 Other Strategies to Increase Power

If the power analysis suggests observing more subjects than is practical, you can consider alternative methods to create a high-powered study with fewer subjects. These methods involve increasing the size of the treatment effect or changing the alpha level of the statistical test that you plan to use. We will consider effect size first. Effect size, as you recall, is defined as the mean difference between the treatments, divided by the standard deviation of the scores: d = MD/SD. Effect size can be increased either by increasing MD or by decreasing SD.

Increasing power by increasing MD. Let's reconsider the study on the effect of training on the SAT. Imagine the training program is a 2-hour seminar covering strategies for answering multiple-choice questions and working examples of the types of questions that can be expected on the test. Compare this to a training program involving 1 hour a day for a full year, in which the content of the test is studied extensively. This yearlong program should be more effective than the 2-hour session. Consequently, the effect size for the yearlong program should be larger and the power to detect the more effective program greater.

The lesson here is that by selecting a treatment with a high likelihood of being effective and comparing it to a treatment expected to be ineffective, for example, a control group, you can expect a larger effect, which will require fewer subjects to achieve adequate power.

> To increase power, plan your study to compare treatments with markedly different effects.

Increasing power by reducing the standard deviation. The SAT training example, discussed above, used an unselected group of high school seniors expected to have a SD of 100 on the SAT verbal exam. By systematically selecting subjects, it is possible to reduce this SD and thereby increase the effect size and the power. Since SAT scores are correlated with grades, subjects could be selected who have average grades; say, a C average. For this select group, the variability of SAT scores should be lower than 100, since there would be fewer high scores and fewer low scores.

Doing the study with these students would increase power through reducing SD.

> The more homogeneous the subjects in a study, the smaller will be the SD and the greater the power to detect a difference between treatments.

There is a disadvantage to improving power by systematically selecting subjects, though. By restricting the subjects, you reduce your ability to generalize the results. In the new version of the SAT study, for example, you would find out nothing about how the training would affect students with averages above or below C.

> The standard deviation also can be reduced by using a more reliable measure for the dependent variable, if one is available, and by tightening the controls in the study.

If you are using a "home-made" measure or rating scale, you might look instead for a published measure with established validity and reliability. Controls could be improved in any number of ways depending on the specifics of the study. It might be possible to increase control by reducing distracting or extraneous events during the experiment or by using uniform procedures, for example, conducting the study in a quiet room, free of interruptions, and tape recording the instructions so that they are the same for all participants.

Increasing power by increasing the alpha level or using one-tailed tests. It is traditional to set alpha equal to .05, guaranteeing a 5% probability of a Type I error. If alpha is set at a higher value, say, .10, the power of the test will be increased. At first glance, this seems appealing; however, it usually is not a good idea because this change also increases the probability of making a Type I error to 10%.

Researchers customarily test the null hypothesis against the alternative hypothesis that the experimental conditions have different effects (direction unspecified); the researcher rejects the null hypothesis if the mean of one condition, say, Al, is sufficiently *greater than* the mean of the other condition, A2, or if the mean of Al is sufficiently *lower than* the mean of A2. Here the statistical test is called *two-tailed* because the researcher rejects the null hypothesis if either of these outcomes occurs. The power chart in Figure 10 is based on a two-tailed test with alpha equal to .05.

However, a researcher might decide instead to test the more specific hypothesis that one of the conditions, say, Al, has a greater mean than the other, A2. Here the null hypothesis would be rejected only for one outcome, when the mean of Al is sufficiently *greater than* the mean of A2, a *one-tailed test.* One-tailed tests, which are more powerful than two-tailed tests, are appropriate, for example, when comparing new treatments with placebo treatments or when testing a prediction deduced from a theory.

6.6 Statistical Conclusion Validity

Virtually every modern experiment that employs random assignment of subjects to conditions also uses statistical tests. Even though the results of statistical tests give the most accurate conclusions possible, these results may be in error. Cook and Campbell (1979) discuss the accuracy of conclusions based on statistical tests as *statistical conclusion validity:*

> *Statistical conclusion validity* refers to the validity of the conclusion of a statistical test.

As you know, statistical tests are subject to two types of error, Types I and II. Cook and Campbell's analysis focuses on the circumstances leading to these errors. Here, we will consider four major threats to valid inference they discuss: low statistical power, violated assumptions of statistical tests, the error rate problem, and experimental instability.

6.6.1 Low Statistical Power

If the power of a statistical test is low, there is a high risk of overlooking the effect of an independent variable. After studying experimental designs in psychology, Cohen (1988) concluded that too many research studies are done with low power. In the previous section, we discussed the steps that researchers can take to increase the power of their studies.

6.6.2 Violation of Assumptions

In order to calculate the *p* value associated with a statistical test, assumptions must be made about the nature of the observations. The *t* test, for example, assumes that the observations are samples from a large set of scores and that the observations in this large set have a normal probability distribution. If this assumption is incorrect, the p value may be inaccurate and the conclusion based on p invalid. The impact of the violation of assumptions is a technical problem that is studied in statistics. To avoid such problems, researchers should be familiar with the assumptions of the tests they use and be confident that any violations of these assumptions, if present, will not affect their conclusions.

6.6.3 Error Rate Problem

As we have discussed, the Type I error rate of statistical tests is controlled by the experimenter. It is customarily set at 5%, meaning that in 5 out of 100 experiments when the null hypothesis is true, the experimenter will incorrectly reject this hypothesis. This is the case if the experimenter conducts only *one* test; when multiple tests are done, the error rate increases. If an experimenter conducts, say, 100 tests, the probability of making *at least one* Type I error can be as high as .99. If you do enough tests, you are virtually certain to make a Type I error, that is, falsely claiming statistical significance.

The increased error rates associated with multiple tests can make the results of research difficult to interpret. For example, a large-scale study done in Sweden reported a statistically significant risk of disease associated with living close to power lines (Feychting & Ahlbom, 1993). However, this conclusion now is in doubt because the researchers conducted hundreds of statistical tests but published only selected results. Although the error rate problem with multiple tests has been recognized for over 30 years (see Ryan, 1959), there still is no satisfactory solution.

6.6.4 Instability

Instability is the threat that circumstances in the experimental situation, other than those associated with the independent variable, may affect subjects' scores on the dependent variable. Such circumstances would include unreliable measures, unwanted variation in treatments, unexpected events during the experiment (e.g., equipment malfunction), and differences among the subjects in characteristics that influence their behavior in the study. All these factors can increase the variability of subjects' scores on the dependent variable, consequently reducing the power of the study and its statistical conclusion validity. The threat of instability can be reduced by using reliable measures and uniform treatments, and by selecting more homogeneous subjects.

6.7 A Comment

With this discussion of power and statistical conclusion validity, we complete our presentation of the basic technical aspects of psychological research. We have seen how Mill's methods are applied and how statistical controls and randomization are used in overcoming the problems of uncontrolled variables. We also presented the basic designs used in experimental and correlational research. These technical issues, however, do not give a complete picture of the concerns involved in conducting research. We have yet to discuss the ethics of research, the moral rights and wrongs that must be considered in studying animals and people.

The ethical codes that we use in psychological research developed in a different manner from advances in technical methods. Methodological advances typically have come from single scientists trying to solve problems in their own research. Correlation was invented by Galton to study heredity; Fisher developed randomization to improve agricultural research. As we discuss in the next chapter, ethical codes and procedures came from committees reacting to the abuse of subjects by researchers. These committees worked to define the basic rights of subjects in research and to develop effective procedures to help researchers safeguard them.

6.8 Key Terms

Measure of error

Randomized blocks design

Replication in the randomized blocks design

Random versus systematic assignment

Null Hypothesis

Significance probability

Alpha level

Statistical significance

Level of significance

Randomization test

t test

Completely randomized design

Latin square design

Yoked control designs

Type I and Type II errors

Power

Standardized mean difference, d

Pilot

6.9 KEY PEOPLE

Ronald A. Fisher

William Gosset, pen name “Student”

Mark Wolraich et al.

Richard Held and Alan Hein

A. J. Rush et al.

Irene Elkin et al.

Jersey Neyman and Egon Pearson

Jacob Cohen

6.10 Review Questions

1. Why did Fisher think that past records of wheat growth on plots A1 and A2 would help him to interpret the results of his experiment on the effect of fertilizer?

2. How did Fisher use these past records to determine a measure of error?

3. Describe how treatments are randomly assigned to plots in the randomized blocks design.

4. What are two advantages of random assignment over systematic assignment?

5. State the rule for reaching a conclusion about the null hypothesis based on the value of *p*.

6. Explain why rejecting the null hypothesis does not mean that the results of a study prove that the null hypothesis is false.

7. Describe how treatments are assigned to plots in a Latin square design.

8. Describe how participants can be grouped into blocks in a psychological experiment using a randomized blocks design.

9. Describe a common application of the Latin square design in psychology.

10. Explain how the yoked control design used by Held and Hein controlled for the visual stimulation the kittens received.

11. How is the probability of a Type I error controlled?

12. What factors in research affect the probability of a Type II error?

13. Use the power chart to determine the power of an experiment with d = 0.50 and 50 subjects in each group.

14. Present four strategies for increasing the power of an experiment.

15. Identify four major threats to drawing valid conclusions with a statistical test.