CONTENTS

8 FACTORIAL DESIGNS AND INTERACTIONS

> *In expositions of the scientific use of experimentation it is frequent to find an excessive stress laid on the importance of varying the essential conditions only one at a time .... this simple formula is not very helpful.*
>
> *R. A. FISHER*

If you pick up a can of Diet Coke and read the label, you will notice a mysterious tongue-twister:

> Phenylketonurics: Contains Phenylalanine.

This message is a warning to phenylketonurics, people suffering from phenylketonuria (PKU), a genetic disorder that prevents the normal metabolism of the amino acid phenylalanine. When this acid, an ingredient in aspartame (the sweetener in Diet Coke), is not metabolized properly, it builds to toxic levels, causing brain damage. Today all newborns are tested for PKU. If the test is positive, the baby immediately is put on a special diet low in phenylalanine. Without this diet, PKU babies (one in about every 15,000 births) would end up brain damaged and institutionalized with the diagnosis of inherited mental retardation.

The discovery of the dietary "cure" for PKU was a fortunate result of scientific advances in how scientists think about development. Most researchers today no longer try to measure the degree of inheritance of traits, as Galton did; geneticists now assume that both heredity and environment *interact* in setting the course of development, as the following quote from Richard Lewontin illustrates:

> *To predict what an organism will be like at some future moment, it is not sufficient to know what it is like now, nor is it enough to describe the environment through which the organism is about to pass. We must know both. (Lewontin, 1982, p. 17)*

PKU is a perfect example. Most people thrive on a diet containing phenylalanine; only the rare people with PKU suffer brain damage because of it.

To develop a cure for PKU, scientists had to be aware of the possibility of *interactions* in human development.

> An *interaction* occurs when the effect of one variable, *A*, on another variable, *X*, depends on a third variable, B.

Table 1 illustrates the interaction between diet, variable A, and genetic type, variable B, in determining brain functioning, variable X.

In everyday language, if the effect of one treatment *depends on* something else, this is an interaction. If you were to ask, for example, "What are the effects on health of drinking alcohol?", a good answer would be that it depends on your age, your sex, and how much you drink. For a woman older than 50, light drinking reduces the risk of death; but if the woman is between 30 and 50, light drinking increases it (Fuchs et al., 1995). Age and drinking *interact* in determining risks to health, in other words. (Heavy drinking at any age increases the risk of death!)

Given the complexity of organisms, it is not surprising that scientists are discovering that interactions are the norm rather than the exception. Few treatments affect every person or animal in the same way. In fact, it is difficult to

TABLE 1 INTERACTION OF DIET AND GENETIC TYPE IN DETERMINING WHO SUFFERS BRAIN DAMAGE FROM PKU

		Diet, Variable A	
		With phenylalanine	Without phenylalanine
Genetic Type Variable B	PKU	Brain Damage	Normal
	No PKU	Normal	Normal

think of treatments that do not interact with other conditions. Should you take aspirin for headaches? It depends on your age. Children can get Reye's syndrome, an often fatal disorder, from taking aspirin; but

aspirin is fine for adults and may even reduce the risk of heart disease. How about penicillin for a strep throat? Again, it depends. Some people have a severe allergic reaction to penicillin. If you become clinically depressed, should you take Prozac? Once more, it depends. If the depression is bipolar, you might be better off on lithium.

Because experience has shown that interactions occur so frequently, researchers now deliberately hunt for them. Numerous studies have been done to find out whether two common treatments for psychological disorders, drug therapy and psychotherapy, combine additively or interact in affecting patients' behaviors. If the drugs alone result in a certain amount of improvement, A, and psychotherapy also results in a fixed improvement, B, when the patient receives both, what will the outcome be? Will the effects add so that-the patient improves by $A + B$, or will the treatments interact to produce a super treatment (a greater improvement than $A + B$), or no effects at all (if the effects cancel each other out)? The answer is critical to finding the best possible therapy.

Psychologists also investigate whether particular treatments interact with patient types. Such research can yield important information on the generality of treatments and the causes of disorders. For example, Stewart, Quitkin, Terman, and Terman (1990) investigated whether two types of depression share the same underlying cause by examining the interaction between treatment and type of depression.

Stewart and his colleagues knew that seasonal affective disorder (SAD), a winter depression, could be treated successfully by exposing patients to bright artificial light during the winter months (the light makes up for the reduced natural sunlight at that time). But they wondered whether light therapy also would help atypical depression, a mood disorder that shares symptoms in common with SAD. They reasoned that if light treatment works as well for atypical depression as it does for SAD, the two disorders actually might be variants of the same underlying problem. When they did the study, they found that SAD responded well to light treatment but atypical depression did not. Their finding, an interaction between treatment and type of depression, supported the standard classification of SAD and atypical depression as separate disorders calling for different treatments.

Experiments such as this one, which test for interactions, must include at least two independent variables and one dependent variable. Until the 1920s, when R. A. Fisher introduced the factorial design, there were no experimental designs for testing interactions. The available research designs before Fisher were modeled exclusively on Mill's method of difference, which states that experimental conditions should be varied only one at a time. This requirement prevented researchers from studying interactions. Fisher's factorial design, to which we now turn, is one of the most commonly used research designs in psychology today.

Fisher (1926) introduced the factorial design by discussing an experiment testing the effects of fertilizers on the yield of winter oats. We will use a psychological example, the effect of drugs on memory, instead of his example from agriculture. Our discussion will follow the logic of this design presented by Joan Fisher Box (1978), Fisher's daughter and his biographer.

8.1 The Factorial Design

Let's begin by imagining that an experimenter is interested in testing the effects of two drugs, A and B, on memory. Some common social drugs that affect memory are alcohol, caffeine, and nicotine (Kerr, Sherwood, & Hindmarch, 1991). In the experiment, participants would learn a list of nonsense syllables, then, after a period of time, receive the drug treatment, then try to recall the previously learned materials.

The two independent variables would be (1) the presence versus absence of Drug A (e.g., caffeine) and (2) the presence versus absence of Drug B (e.g., nicotine). The dependent variable would be a measure of the amount of material recalled. Since there are two independent variables and each variable has two different values or *levels* (presence vs. absence), there are 2 x 2, or four, different treatments in the study (see Table 2).

> In a factorial design, the total number of treatments is equal to the product of the number of levels of each of the independent variables.

A list of the treatments in this experiment can be generated by multiplication. If we call the levels of independent variable *A*, *a1* and *a2*, and the levels of independent variable *B*, *b1* and *b2*, then the product of *(a1 + a2)* times *(b1 + b2)* gives the full set of treatments:

(a1 + a2) (b1 + b2) = a1b1 + a1b2 + a2b1 + a2b2

where a1b1 stands for treatment a1 and b1 given together.

TABLE 2 THE FOUR TREATMENT CONDITIONS IN THE 2x2 FACTORIAL DESIGN STUDYING THE INTERACTION OF DRUGS A AND B

	Drug A	
Drug B	Absent	Present
Absent	Placebo	Only Drug A
Present	Only Drug B	Drugs A & B

The factorial design gets its name from this process of multiplying to yield the experimental treatments. If you recall from algebra, an equation like $X^2 + 3X + 2$ can be factored into the product of two terms involving X. These terms, (X + 1) and (X + 2) in this case, are the factors of the equation. Similarly, the terms *(a1 + a2)* and *(b1 + b2)* are the factors of the experimental design, since they can be multiplied together to give the full set of possible treatments. The terms *factor* and *independent variable* are used interchangeably.

> *A factorial design* is an experimental design with two or more independent variables, in which the complete set of treatments or conditions is generated by multiplying together the levels of the independent variables.

Factorial designs are described by giving the number of levels on each factor. The memory study is called a 2 x 2 ("two by two") factorial design, because each of the factors has two levels. If one factor had 4

levels and the other factor 3 levels, it would be called a 4 x 3 ("four by three") factorial design.

If the full set of treatments is not used, the experiment does not have a factorial design. If the memory study had only three conditions, for example, Drug A, Drug B, and Placebo, it would not be a factorial design, even though each level of each independent variable would be present.

8.1.1 Assigning Subjects to Treatments

There are two general procedures for assigning the subjects to the four different treatments of our 2x2 memory study. In a within-subjects design, each subject would receive all four treatments. In a between-subjects design, each subject receives only one treatment, so different subjects would be used in each condition.

Which of these two designs to use would depend upon the specifics of the study. Within-subjects designs have the advantage of controlling for individual differences among the subjects, because subjects' characteristics are constant across the treatments. In addition, this design requires a fraction of the subjects needed for a between-subjects design. In the memory study, if we wanted to have, say, five observations in each treatment group, the between-subjects design would require 5 x 4 = 20 subjects, but the within-subjects design would require only 5 subjects. The problems of the within-subjects design result from the repeated measurement of the same subjects. When each subject participates in all the treatments, fatigue and practice effects can result. If the treatments are given in different orders, interaction effects are possible (e.g., drug A may have a particular effect on memory when preceded by drug B but not when preceded by a placebo). Procedures for controlling for these problems are discussed later in the chapter. The between-subjects design avoids the problems of repeated measures by having each subject receive only one treatment.

We will explain the logic of the factorial design using the 2x2 memory study with a between-subjects design. Let's assume that 20 people are randomly assigned to the four drug treatment conditions, with 5 people in each of the four conditions: Placebo (P), Drug A (A), Drug B (B), and Drugs A and B (AB). The last treatment, giving participants

both drugs simultaneously, would not be included in an experiment varying the treatments one at a time. In fact, at first glance this treatment seems to make it impossible to untangle the effects of the two drugs. How is it possible to figure out the influence of each drug when both are given to the same subjects? Fisher had an ingenious answer to this question that, oddly enough, is based on the logic of varying one thing at a time!

8.1.2 Comparisons between Conditions

The first step in analyzing the results of a factorial experiment is to calculate the mean (average) values of the dependent variable for the different experimental conditions. Table 3 shows the individual subject scores and the means for each treatment in the memory study. The scores are the results of the memory test given to the subjects after taking the drugs—the higher the score the better the recall.

8.1.3 The Interaction between Factors A and B

In Table 3 there are two comparisons between the means of the conditions that provide information about the effects of Drug A. Each of these comparisons is based on the logic of the method of difference; as required by the method, only one condition is varied for each comparison.

1. M_A-M_P

   The mean of condition A, M_A, can be compared to the mean of condition P, Mp ; the difference between these means gives the advantage of Drug A over the placebo. Getting the means from Table 3, M_A - Mp= 15 -10 = 5. The observed effect of Drug A here is to increase recall by 5 points.

2. M_{AB}-M_B

The mean of condition AB, M_{AB}, also can be compared to the mean of condition B, M_B. The difference between these means gives the advantage of giving Drug A to subjects who also are receiving

TABLE 3 SCORES ON RECALL AND MEAN RECALL SCORES FOR EACH TREATMENT CONDITION IN THE 2x2 FACTORIAL DESIGN

		Drug A	
		Absent	Present
Drug B	Absent	P	A
		S 1: 10	S 6: 10
		S 2: 7	S 7: 7
		S 3: 9	S 8: 9
		S 4: 9	S 9: 9
		S 5: 15	S 10: 15
		M_P = 10	M_A = 15
	Present	B	AB
		S 11: 10	S 16: 10
		S 12: 7	S 17: 7
		S 13: 9	S 18: 9
		S 14: 9	S 19: 9
		S 15: 15	S 20: 15
		M_B = 20	M_{AB} = 33

Drug B. The result is $M_{AB} - M_B = 33 - 20 = 13$. The effect of A *in the presence of B* is to increase recall by 13 points.

Both comparisons show an increase in recall when the participants take Drug, A, but the advantage of A is greater when B is present, 13 points, than when B is absent, 5 points. This result is evidence of an interaction between A and B. If A had the same effect in the presence or absence of B, there would be no evidence of an interaction.

To evaluate the possibility of an interaction between Factors A and B, comparisons must be made of the effects of A at different levels of B.

In our example, the effect of A differs by 13 - 5 = 8 points depending on whether B is present or absent. This difference provides a

numerical measure of the strength of the interaction; the greater this number, the more evidence of an interaction.

The null hypothesis, that the interaction value is equal to zero, is tested by computing the significance probability, p, which is the probability if the null hypothesis is true of getting the observed value, or one even greater. As in every statistical test, if p is less than or equal to the alpha level chosen for the test (usually alpha = .05), the null hypothesis is rejected. In the example, the statistical test results in $p < .05$, so there is a significant interaction.

This statistical test for the interaction, devised by Fisher, now is called the F (for Fisher) test to recognize his work. The computations for the F test are explained in statistics texts and handbooks of experimental design (see Winer, 1991, or Kirk, 1982).

8.1.4 Main Effect of Factor A

In some experiments, the researcher may be interested in the average effect of an independent variable. In our example, Drug A increases recall by 5 points when B is absent and 13 points when B is present; so in this case, the average or *main effect* of A is (13 + 5)/2 = 9 points.

The *main effect* of Factor A is determined by computing the effect of A at each level of B and averaging these values.

The main effect of a factor, like the effect of an interaction, can be tested for significance by an F test. The null hypothesis in this case is that the main effect

To evaluate the possibility of an interaction between Factors A and B, comparisons must be made of the effects of A at different levels of B.

In our example, the effect of A differs by 13 - 5 = 8 points depending on whether B is present or absent. This difference provides a numerical measure of the strength of the interaction; the greater this number, the more evidence of an interaction.

The null hypothesis, that the interaction value is equal to zero, is tested by computing the significance probability, p, which is the probability if the null hypothesis is true of getting the observed value, or one even greater. As in every statistical test, if p is less than or equal to the alpha level chosen for the test (usually alpha = .05), the

null hypothesis is rejected. In the example, the statistical test results in $p < .05$, so there is a significant interaction.

8.1.5 Main Effect of Factor B

In the 2x2 factorial design, the analysis that is done for Factor A is repeated for Factor B. Again, two comparisons are needed to judge the effect of Factor B:

1. M_B-M_P

The mean of condition B, MB , first is compared to the mean of condition P, M . The difference gives the advantage of Drug B over the placebo. Getting the means from Table 3, M_B - M = 20 - 10 = 10 points.

2. M_{AB}-M_A

The mean of condition AB, M_{AB}, then is compared to the mean of condition A, M_A. The difference gives the advantage of Drug B for *subjects who also receive Drug A*. The result is M_{AB} - M_A = 33 - 15 = 18 points.

The main effect of B is calculated by averaging the results of these two comparisons: (18 + 10)/2 = 14. Averaged over levels of A, the main effect of B is 14 points. The F test of this main effect is significant at $p < .05$.

We also can use the observed effects of B at the different levels of A to determine if there is an interaction between B and A. The effect of B when A is absent is 10 points; when A is present, it is 18 points. So the effect of B differs by 18 -10 = 8 points, depending on the level of A. Notice that this is the same value, 8, that we got when we calculated the interaction of A and B before. This is no coincidence.

The evidence for an interaction between Factors B and A is always the same as the evidence for an interaction between A and B. Consequently, there is only one F test for the interaction.

TABLE 4 ANALYSIS OF VARIANCE SUMMARY TABLE FOR RECALL SCORES

Source	*df*	*F*
Drug A	1	38.12*
Drug B	1	92.24*
A x B	1	7.53*
Error	16	(10.63)

The value in parentheses is the mean squared error.

* $p < .05$.

The context of the experiment usually will favor one or the other way of stating the interaction—either that the effects of A depend on B, or that the effects of B depend on A.

8.1.6 The Analysis of Variance

The complete analysis of the two factor design, called the *analysis of variance*, includes three F tests—one test for each main effect and one test for the Interaction. These tests are presented in an analysis of variance summary table. A standard format for such tables is shown m Table 4. The first column, labeled Source, lists the names of the main effects, the interaction, and the error term. The error term is used in computing the F tests, Its value (shown in parentheses) is a measure of the extent to which differences among the scores on the dependent variable are due to uncontrolled variables.

The second column shows the degrees of freedom, 'if, associated with the main effects, the interaction, and the error. These numbers are based on the size of the study. For each main effect, if is equal to one less than the number of levels of that factor. In our example, each factor has two levels; so, *df* = 2 - 1 = 1. The *df* of the interaction is the product of the *dfs* of the two main effects, 1 x 1 = 1, The *df* for error depends upon the number of subjects and the number of treatments. In the example, df is equal to the total number of subjects (20) minus the number of treatments (4); df error = 20 — 4 = 16.

The third column shows the value of the test statistic, F, for each statistical test. The larger the value of F, the smaller the value of the significance probability, h". Fisher published tables of the critical values of F for different values of alpha. The table, available in most statistics texts, shows that each F test in our example is statistically significant at $p < .05$.

8.2 Variations in Factorial Designs

In Chapter 5, we discussed how Fisher's agricultural experiments with one independent variable translated into experiments in psychology. We considered three designs: (1) between-subjects designs in which subjects are assigned to the conditions completely at random (the completely randomized design), (2) the between-subjects design in which subjects are blacked before being randomly assigned to conditions (the randomized blocks designs), and (3) the within-subjects designs in which each subject is observed in each condition of the study (the repeated measures Latin square design). Each independent variable in a factorial design can be based on any one of these three designs, Subjects in a factorial design can be

- Assigned completely at random to the levels of factor,
- Be matched into groups (blocks) and then randomly assigned to levels of the factor (randomized blocks) or
- Observed at each level of the factor (repeated measures)

There is one additional way that subjects can be assigned to the levels of a factor, one that we did not discuss in Chapter 6; namely

subjects can be systematically placed in, not randomly assigned to particular levels of a factor. Systematic assignment is based on characteristics of the subjects, like gender, age, or personality type. An independent variable based on systematic assignment is called a subject factor.

Two between-subjects factors—completely randomized (no matching) and randomized blocks (matching)—were discussed In Chapter 6. Applying these methods in factorial designs raises no new Issues. The other two methods— repeated measures and subject factors—do have special problems that we Will discuss after we look at some examples of factorial experiments with between subjects factors.

8.3 BETWEEN-SUBJECTS FACTORS

Factors with complete randomization are routine in psychological research. In studies using this method, subjects first are selected who are similar on variables that the researcher suspects might Influence the outcome of the treatments (e.g., age, severity of a disorder); the subjects then are randomly assigned to the treatment groups. Often repeated measurements are made on the dependent variable before and after the treatment.

Th1S design has much to recommend it. There are no major threats to its internal validity and it ts easy to use. The randomization can be preplanned and subjects can be assigned to the conditions when they volunteer for the study. Randomization also allows the investigator to calculate a measure of error due to uncontrolled variables,

Elkin et al, (1989) used this design m a National Institute of Mental Health sponsored large-scale experiment comparing drug therapy and psychotherapy for depression. The patients in the study were randomly assigned to either (1) drug therapy, (2) cognitive psychotherapy, (3) interpersonal psychotherapy, or (4) a placebo drug treatment. Their depressive symptoms were assessed before, during, and at the end of therapy, as well as at 6-, 12-, and 18-month intervals following termination. In this evaluation study, patients were randomly assigned to the levels of the first factor, the type of treatment, and the degree of their depression was measured at each level of the second factor, the time of measurement. The first factor is a between-subjects factor because different patients are observed at each of its levels; the second is a within-subjects factor because the same subjects are observed at every level. A design having both between and within factors, like this one, is called a mixed design.

Randomization also can be used for both factors of an experimental design. This was done by Sigall and Ostrove (1975) who investigated the role of a convicted felon's appearance on the sentence she was given in a criminal trial. The subjects, who played the role of jurists in the study, read a description of a crime committed by a woman whose photograph was attached to the description. In facU the photographs and the felony described to the subjects varied. The photo was either of an attractive or an unattractive woman (Factor A) and the crime either was a swindle or a burglary (Factor B). The participants were

randomly assigned to one of the four possible conditions (2 x 2) and were asked to decide how many years in prison would be a suitable punishment for the crime. The results revealed an interaction between the type of crime and the attractiveness of the felon. The attractive burglar was given a shorter sentence than the other three combinations.

Between-subjects factors with complete randomization were ideal for this problem, since the alternative, using within-subjects factors, would require subjects to sentence all four cases. If this design were used, most likely the nature of the manipulation, varying attractiveness and the type of crime, would become apparent to the subjects, possibly affecting the outcome of the siudy.

Randomization with prior matching is a good alternative to complete randomization. Subjects can be matched on important variables and then randomly assigned ta levels of a factor. Matching is an excellent method for reducing error due to uncontrolled individual differences between subjects. Azrin and Peterson (1990) used matching in this way to evaluate a behavioral treatment for Tourette syndrome, a disorder characterized by involuntary motor tics and embarrassing verbal outbursts.

The experimenters wanted to evaluate the effects on patients of receiving a behavioral treatment for their disorder. However, only 10 subjects with Tourette syndrome were available for the study, and they had very different ages, varying degrees of severity of their symptoms, and they differed on whether they took medication for the problem. Random assignment to the two planned experimental groups (treatment vs. no treatment) would have been unwise with such a small, diverse subject pool, because the groups might have ended up being quite different on some combination of these variables, as an alternative, the authors used a randomized blocks design.

Five pairs of subjects were formed, with both members of a pair matched so that they were almost the same age and had symptoms of about the same severity Both were matched on whether they were on or off medication. Then one member of each pair was randomly assigned to the habit reversal treatment and the other to a waiting list to be treated at the end of the study. Type of treatment is a between-

subjects factor because different subjects are assigned to the levels of the factor. The frequency of the subjects' tics was observed before and after the treatment (or waiting period), within-subjects factor based on repeated measures. The results showed an astounding 92% reduction of tics for the behavioral treatment, far better than the reduction rate with medication alone. These results raise hope of a major breakthrough in the treatment of this devastating disorder.

This study used matching to control for differences among the patients on age, severity of symptoms, and medication. The investigators could have used a different strategy based on the factorial design if they had had more subjects Using this new design, the variables that Azrin and Peterson controlled through matching could have been introduced as separate factors. Such a factorial design would permit statistical tests of the main effects and interactions of these variables. The Drug x Treatment interaction, for example, would test whether the behavioral treatment worked better or worse when the patient was on medication. Unfortunately, introducing new factors requires large numbers of subjects, so this design is not useful for studying rare conditions, like Tourette syndrome.

8.3.1 Within-Subjects Factors

Both of the therapy evaluation experiments we have discussed had repeated measures on one factor. There are no special problems with within-subjects factors when they are used in this way, to record changes in subjects aver time, Problems do arise, however, when within-subjects factors are used to evaluate different treatments. In such cases, the order of presenting the treatments becomes an issue, An experiment by Hall and Kataria (1992), which evaluated different treatments for attention deficit hyperactive disorder (ADHD), a condition characterized by impulsive, overactive, and inattentive behavior, illustrates one solution to this problem.

The children in the study were randomly assigned to either a behavioral, cognitive, or control (inactive) treatment. Repeated measures were taken on each child under medication (Ritalin) and with no medication Hall and Kataria found a significant interaction between the medication and the psychological treatment for the

children's performance on a delayed response task Medication combined with cognitive treatment resulted in the best performance.

The delayed response task was given to the children twice, when they were taking Ritalin and when they were not. When the same task is repeated, as it was in this study, there is always the danger of order effects; that is, repeating the task may affect the results, either through practice, fatigue, or boredom, So, in this case, it would not be desirable to give every child the drug first, followed by the no drug condition.

For this reason, the experimenters decided to counterbalance order by giving half the children the Ritalin first, and giving it to the rest of the children second, With this procedure, any order effects are balanced, since equal numbers of subjects receive the treatments in each possible order. The Latin square design discussed in Chapter 6 is based on this logic,

Counterbalancing effectively controls for changes that take place m subjects during an experiment, like fatigue or boredom, but does not guard against interactions among the treatments. If, for example, treatment B is very effective when it follows A but ineffective otherwise, the results for treatment B will depend on how many times B follows A. If B never follows A, the researcher would erroneously conclude that B is an ineffective treatment. For this reason, counterbalancing should not be used when investigators suspect the possibility of interactions involving order.

The logic of the factorial design offers an excellent alternative to counterbalancing for dealing with order effects. Order can be introduced as a separate factor in the design. Using this strategy, each possible order of the treatments would be a level of the factor to which subjects would be randomly assigned. This procedure would allow the experimenter to test for order effects as well as interactions between order effects and the other factors in the design.

Although this may be the best way to control for order effects, it is not problem free. With several treatments, the number of levels of the order factor becomes excessively large, requiring large numbers of subjects. With five treatments, for example, there are 120 possible orders; assigning subjects to each of these orders would require many

subjects. With only two or three treatments, having two and six possible orders respectively, this technique would be worth considering.

The simplest procedure for dealing with order effects is randomization, a popular method when an experiment involves many treatments or tasks, Using randomization, subjects are assigned to the treatments in an order selected at random. With four treatments, for example, there are 24 possible orders. Each sub)ect would be assigned, by chance, to one of these. This method of assigning subjects to orders deals with order effects in the same way that randomly assigning subjects to groups handles uncontrolled individual differences. It avoids any systematic bias. Of course, the possibility remains that the orders that are selected will favor some treatments.

Regardless of which strategy is used to deal with order effects, the experimenter's job is to set the procedures of the study to minimize them. This might be done by introducing a break between treatments (e.g., giving the treatments on different days) or by designing the treatments with a view tc minimizing fatigue, boredom, and practice effects. In the Ritalin study, for example, a 24-hour period during which the subjects were off the drug preceded the no-drug condition.

8.3.2 Subject Factors

Subject factors, as you remember, are factors In a study that are based on characteristics of the subjects. The levels of a subject factor are classifications of the subjects on such variables as personality, age, or gender. Because different subjects are assigned to the levels of subject factors, they are between subject factors.

Such factors are included in psychological research for reasons. They might be:

- The primary focus of the study. In personality research, for example, people are classified into types, such as the Type A or Type B personality, and compared to see whether their behaviors differ. The competitive, achievement-oriented Type A person has a greater risk of coronary heart disease than the more relaxed, mellow Type B personr far example (Jenkins Zyzanski, & Rosenman, 1979).

- A substitute for factors that cannot be manipulated for practical or ethical reasons. It would be unethical, for instance, to subject people to high levels of stress to observe changes in their immune systems. However, people can be measured on stress and then studied, as Cohen et al. did (Cohen et al., 1991; see Chapter 4). These researchers gave a viral challenge, nose drops loaded with common cold virus, to subjects who were either high or low on measured stress. They found that a higher percentage of the high stressed subjects caught the cold,
- Used for assessing external validity. If a main effect or interaction between factors can be shown to hold for subjects differing on characteristics like age or gender, its generality is assured. This is commonly done in survey research in which sample sizes are large enough to study several factors at once.
- Used as a first step before beginning a more elaborate study. Subject factors can be used to establish differences between participants, which then can be examined more fully in a more extensive study. For example, research has shown that there is a sex difference in alcohol metabolism. When alcohol consumption is proportional to body weight, less alcohol is found in men's bloodstreams than in women's. This finding, which confirms the folklore that men "hold their liquor" better than women, was a necessary preliminary to later research testing hypotheses about the basis for this difference. We now know that men have an enzyme in their stomachs that breaks down alcohol before it reaches the bloodstream; because women have less of this enzyme, more alcohol reaches their bloodstream (Frezza et al., 1990). Men apparently hold their liquor in their stomachs.

Subjects, of course, cannot be assigned randomly to the levels of a subject factor, and sometimes the differences between levels of such a factor are not unitary. Consider "gender," for example. Subjects are not randomly assigned to a gender and there is nat a Single difference between men and women. Consequently, if the main effect of a subject variable is significant or if the interaction between a subject variable and another variable is significant, the result often is difficult to interpret. Finding a significant subject factor is the same as finding a

significant correlation between variables, so the same problems arise in interpreting such effects as in interpreting correlations (see the discussion of these problems in Chapter 5)- The following study illustrates the use of a subject factor in a factorial design.

•Calvin and Hobbes C) VVatterson. Dist. by UN VERSAL PRESS SYNDICATE, Reprinted with permission, All rights reserved.

Many adults are convinced that eating sugar negatively affects both the behavior and cognitive abilities of children (see the Cfilvin Hobbes cartoon) but not of adults. But this was not what was found in an elaborate, well-controlled study that varied the amount of sugar in children's diets over nine weeks In Chapter 6, we presented Wolraich et al, 's (1994) study which used a Latin square design to examine the effects on children of diets high in sugar, aspartame and saccharin. They found no effects, adverse or beneficial, of the high sugar diet. After reviewing that study, we were left with the mystery of why sugar has such bad reputation if, in fact; its effects are not negative.

This mystery may have been solved by Jones et al. (1995) in an experiment that used a different method of giving the sugar to the children as well as a different experimental design. Jones and his colleagues studied the short-term effects of having children ingest a large amount of sugar at one time, what they called a standardized large glucose load." Their method was comparable to real-life situations where children eat a large amount of sugar (e.g., at parties, and for Calvin, at breakfast), Unlike Wolraich et al.'s study,

participation was not restricted to children. They loaded adults with glucose as well.

The Jones study also had two independent variables rather than one. Each of the variables had two levels: (1) Sugar Treatment, pre- and post-, and (2) Age, children and adults. The sugar treatment factor involved repeated measures—the participants were observed before and after the glucose load and age was the subject factor. This mixed design had one between-subject factor, Age, and one within-subject factor, Sugar Treatment

The hypothesis of the study was that age and treatment interact in determining subjects' reactions to sugar, Specifically, the authors expected that the reaction to sugar, measured by self-reported symptoms, would be greater for

TABLE 5 MEAN SYMPTOM LEVELS BEFORE AND AFTER SUGAR LOADING FOR CHILDREN AND ADULTS (FROM JONES ET AL., 1995)

	Sugar Treatment	
Age	Pre-sugar	Post-sugar
Children $n = 25$	12.8	22.0
Adults $n = 23$	10.0	13.0

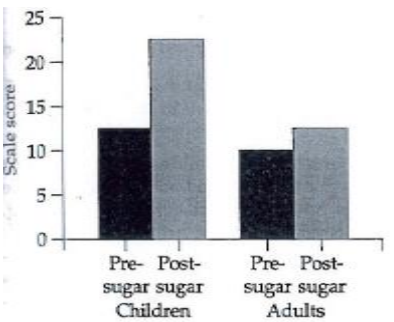

FIGURE 1
Bar chart showing mean symptom scale scores.

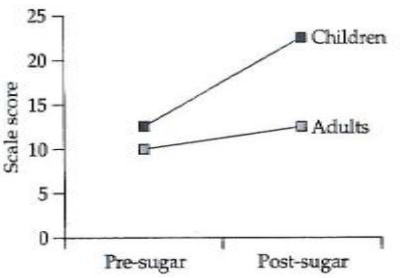

FIGURE 2
Line chart showing mean symptom scale scores.

children than adults. Each symptom (shaky, heart pounding, headaches, feeling weak, anxious, difficulty concentrating, slowed thinking, and feeling sweaty) was rated on a scale from 1 to 7, 1 being

"the symptom is not present at all," and 7, "the symptom is present in the extreme,"

The symptom levels for both children and adults are shown in Table 5. Before the sugar load (administered on a per body weight basis to control for the size differences between children and adults), the symptoms were comparable for both groups. After ingesting the sugar (for the child, an amount equivalent to drinking a 24-ounce bottle of Coke), the symptoms of the children increased more than did those of the adults.

Figures 1 and 2 show these results using two popular types of charts, In Figure 1, a bar chart, each experimental condition is shown as a bar, with the height equal to the mean value of the reported symptoms. Figure 2 is a line chart; each line on this chart connects the pre- and post-treatment means on the symptom scale for one group of subjects. Although investigators choose one of these three methods—the table, bar chart, or line chart—to present their findings in publications, the evidence for interactions can be seen most clearly in the line chart. If its two lines are not parallel, there is some evidence for an interaction.

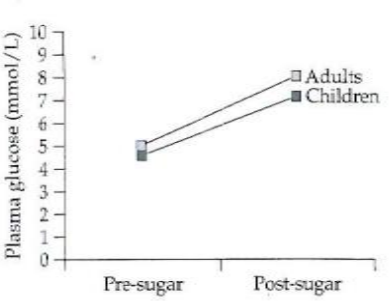

FIGURE 3
Mean glucose levels pre- and post-sugar ingestion for children and adults.

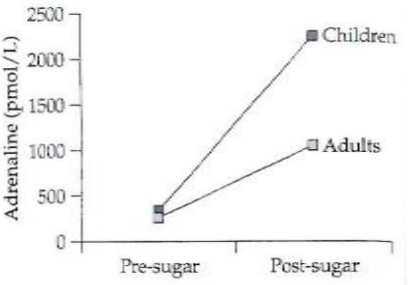

FIGURE 4
Mean adrenaline levels pre- and post-sugar ingestion for children and adults.

Although Jones et al.'s study demonstrated that children and adults react differently to sugar, it still isn't clear why. Children and adults differ in more ways than one, so there also are many possible explanations for this finding. Such problems in interpretation are common for subject factors.

A general strategy, used frequently by researchers to help them later interpret results involving subject factors, is to formulate one or more hypotheses that would account for the expected effects and then to collect additional observations to evaluate them. Plans to collect the additional observations are included in the original design of the study.

Jones et al, anticipated that children and adults would react differently to sugar and speculated that this was because children's adrenal glands are more responsive to blood sugar level than adults'. To test this hypothesis, the participants' blood, which was drawn before and after the sugar load, was analyzed to determine how much glucose, that is, metabolized sugar, and adrenaline it contained. The results of these analyses showed virtually identical levels of glucose in the children and adults following the sugar load (see Figure 3), but different levels of adrenaline, the children's levels increasing much more than the adults (see Figure 4). This result confirmed the experimenters' expectations.

8.3.3 The Number of Factors

The Jones et al. study tested men and women as well as boys and girls. If a different reaction to sugar had been expected for males and females, the design could have introduced sex as a subject factor, a third factor in the design. This additional factor would allow the investigators to test whether sugar loading affects males and females similarly.

Every additional factor that is included in a design increases the number of statistical tests that are calculated, In a three-factor design, there are tests for

FACrORlA1.

three main effects, A, B, and C; tests for three “two-way interactions,” A x B, A x C, and B x C; and a test for the "three way interaction,” A x B x C.

"Two-way" interactions are the same as the interactions we already have discussed. The A x B interaction is tested by examining the effects of Factor A at the different levels of Factor B. The other two-way interactions are defined similarly; the B x C interaction, for

example, is sensitive to whether the effects of Factor B on the dependent variable differ across levels of Factor C.

The three-way interaction, Ax B x C, is a complex idea; it involves four variables, the three independent variables and the dependent variable. A three-way interaction results when a two-way interaction, say A x B, is different for different levels of C,

The A x B interactions at different levels of C are compared to test for the possibility of AX B x C interaction.

The A XB x C interaction could be described as the B x C interaction varying across levels of A; or the A x C interaction varying across levels of B, These are equivalent descriptions, There is only one three-way interaction in a three-factor design,

There is no upper limit to the number of factors in a factorial design, except perhaps human understanding. Trying to interpret a "four-way" interaction, for example, is a serious challenge If the ideas being tested require such complexity, however, the factorial design is unequaled.

8.4 Advantages Of The Factorial Design

In his book *The Design of Experiments*, Fisher (1935) presented three major advantages of the factorial design over traditional experiments that vary one variable at a time. Over 60 years of experience with the factorial design have borne out Fisher's original assessment.

8.4.1 Efficiency

In a factorial design with a given number of subjects, it is possible to test the effects of two (or more) factors with the same precision as a traditional study of equivalent size that tests only one variable. To understand this, imagine that 40 subjects are available for a study In a traditional experiment, 20 subjects would be assigned to Treatment al and 20 to Treatment 02. In like manner, in the 2 x 2 factorialf 20 subjects would get Treatment and 20 would get Treatment 02; but 20 of the subjects also would get Treatment bl and 20 would get Treatment b2. For each factor, it is possible to compare 20 subjects against 20 other subjects. Since this is the same number of subjects

we compared in the traditional study, the factorial design is more efficient, yielding more information from the same number of subjects.

8.4.2 Comprehensiveness

The factorial design also permits tests of interactions. Such tests are not possible in one variable at a time studies. In addition, the precision we saw for main effects also applies to testing interactions.

8.4.3 External Validiti/

In a traditional study, the effects of a single variable are evaluated holding other conditions constant. But this design severely limits the possibility of generalizing from the study, because there is no evidence that the results will replicate across other conditions. In a factorial design, by contrast, the effects of each factor are evaluated at the same time as the other factors are varied, so an assessment of external validity is built into the design. When the interaction is not significant, there is direct evidence that the effects generalize across these conditions. When the interaction is significant, information is gained about the limits of the generalization. In Fisher's words:

> As the factorial arrangement well illustrates, we may by deliberately varying some of the conditions of the experiment, achieve a wider inductive basis for our conclusions, without in any degree impairing their precision. (Fisher, 1935, p, 100)

8.5 The General Linear Model

So far, we have discussed two statistical approaches to the basic problem of uncontrolled variables in research, statistical control and randomization. In Chapter 5, we saw how statistical controls are used in correlational research when random assignment of subjects to conditions is impossible. As you remember, if you can measure subjects on an uncontrolled variable, it IS possible to remove the influence of this uncontrolled variable on the dependent variable by using the mathematics of multiple correlation.

In experimental designs, the random assignment of subjects to experimental conditions avoids systematic bias and allows the

experimenter to calculate a measure of the error due to uncontrolled variables (see Chapter 6). This measure of error is used in statistical tests.

We usually are taught that these two methods, statistical control and randomization should be applied in different types of studies. Researchers use randomization whenever possible; otherwise they are forced to rely on statistical controls. And the methods are taught as different mathematical procedures- In statistical control, the experimenter fits different mathematical models to the data and evaluates their fit. Randomization is followed by statistical tests, such as the analysis of variance, which test differences between the means of experimental groups.

Within the last 25 years, however, a new approach to data analysis that unifies these two traditional approaches has become increasingly popular. *The general linear model* (GLM) is a generalization of multiple correlation that can be used to analyze the results of experiments and correlational research.

GLM not only simplifies the logic of data analysis, since one general method can analyze data from almost any research project, but GL M provides the experimenter with techniques that are not available in the traditional data analysis. The unification provided by GLM permits the best feature of correlational data analysis, statistical controls, to be used in experimental studies, and allows one of the best features of experimental work, the testing of interactions, to be used in correlational research.

GLM is taught today in advanced methods books after the traditional data analysis procedures are presented (e.g., see Winer, 1991, or Kirk, 1982), But the efficiency and scope of CLIM is so great that we would not be surprised if, within the next 23 years, it becomes the dominant method of analysis taught even in introductory texts.

8.6 KEY TERMS

Phenylketonuria

Interaction

Factor

Factorial design

Between-subjects designs vs. within subjects designs

F test

Main effect

Analysis of Variance

Analysis of Variance summary table

Degrees of freedom, df

Between-subjects factor

Within-subjects factor

Counterbalancing

Order effects Mixed design

Subject factor

Bar chart

Line chart

General linear model

8.7 KEY PEOPLE

R. A. Fisher

Joan Fisher Box

T. W. Jones et al.

8.8 REVIEW QUESTIONS

1. Explain what researchers mean by an interaction between variables.

2. Give three examples of interactions from everyday life.

3. Why were Stewart et al. testing for an interaction between type of patient and type of treatment?

4. How can the full set of treatments in a factorial design be generated from the individual factors?

5. Describe a 4 x 3 factorial design.

6. How many subjects would be required in a 3 x 3 between-subjects factorial design to have 10 subjects in each condition of the experiment? How many subjects would be needed if it were a within-subjects design?

7. In a 2 x 2 factorial design, describe how a numerical measure of the strength of the interaction is calculated.

8. How many F tests are there in the analysis of a 2 x 2 factorial design? What effects do they test?

9. What are the degrees of freedom for the main effects and interaction in a 2 x 2 factorial design?

10. Present the four ways of assigning subjects to the levels of an independent variable in a factorial design.

11. Give two examples of factorial designs with random assignment of subjects to at least one factor of the design.

12. How was matching used in the study of the behavioral treatment of Tourette syndrome?

13. What are the problems of using repeated measures in a factorial design?

14. Describe three techniques for dealing with order effects in a within-subjects design.

15. Describe four reasons for including subject factors in a factorial design.

16. Describe the factorial design used in the Jones et al. study. Why is this design called a mixed design? What other studies described in this chapter are mixed designs?

17. What strategy did Jones et al. use to help them interpret the interaction between sugar treatment and age in their experiment?

18. List all the statistical tests possible in a 2 x 2 x 2 factorial design.

19. According to Fisher, what are the three advantages of the factorial design?

20. The general linear model allows a major feature of correlational analysis to be used in experimental research and a major feature of factorial designs to be used in correlational research. What are these two features?