Chapter 12. Experimental Design: One-Way Correlated Samples Design

Advantages and Limitations

In the previous chapter, we included the powerful technique of random assignment in our research design to reduce systematic error (confounding variables). The assignment of participants to groups in a random fashion is one of the best ways to equate the groups on both known and unknown factors prior to administration of the independent variable. However, as we noted, there is no guarantee that they will be equated. To enhance experimental control, you may want to guarantee that one or more variables are equated among your treatment levels, and you may not want to rely on random assignment to establish that equality. Remember that any variable that varies systematically among your treatment levels and is not an independent variable is a confounder that can mask the effect of the independent variable. For example, in the previous chapter we discussed random assignment of children to two groups in a TV violence study. Random assignment, by chance, could result in having more boys in one group and more girls in the other group. If gender of the child is related to the dependent variable (aggressiveness), then we have created a confounding variable that will result in systematic error. As we will see, one advantage of the correlated samples designs discussed in this chapter is the reduction of systematic error between the treatment conditions.

However, the primary advantage of the correlated samples designs is the reduction of random error due to individual differences. Recall that random error creates “noise” that makes it more difficult to detect systematic effects of the independent variable. Reducing the noise enables us to detect smaller differences (systematic variance) between treatments. In terms of statistical formulas, we will see that the denominator of the formulas for our test statistic (*t* or *F*) consists of random error and the numerator consists primarily of the treatment (systematic variance). The larger the random error, the smaller the value is for the test statistic, and the less likely we are to find a treatment effect that is statistically significant.

The three techniques introduced in this chapter are all correlated samples designs. Correlated samples designs do not use random assignment of participants to conditions. Instead, they either test the same research participants under each treatment condition or match different participants on a related factor. Similar to random assignment designs, correlated samples designs can be used with two treatment conditions or more. The three types of correlated samples designs are natural pairs, matched pairs, and repeated measures. We should note that the appropriate statistical test is related to the research design that is used. For example, the *t* test for correlated samples design is calculated differently than that for independent samples design.

Table 12.1 Characteristics of research designs.

Type of Research Design	*Number of IVs*	*Number of Levels of the IV*	*Number of DVs*	*Assignment to Conditions*	*Most Probable Inferential Statistic*
One-way independent samples	1	2 or more	1	Random	*t* test or one-way ANOVA
One-way correlated samples	1	2 or more	1	Natural pairs Matched pairs Repeated measures	*t* test or one-way ANOVA

Natural Pairs

In a natural pairs design, the scores in the groups are paired for some natural reason; an effort is made to match the participants on some natural basis. A good example of this matching would be twin studies. Returning to our TV violence study, research suggests that there is a genetic component to some aspects of personality, including aggressiveness. That is, babies come into this world with temperaments that help shape their developing personalities. Therefore, when we observe levels of aggressive behavior in children in a day-care center, we suspect that part of the explanation for their behavior is their genetic profile. Thus, variability of scores within and between groups is partly due to different participants' having different genetic profiles. This factor contributes to random error and makes it more difficult to detect variability due to the independent variable.

One solution is to eliminate genetic differences between groups by using identical twins. If we place one of the twins in one treatment condition and the other twin in the other treatment condition, we have created a situation in which there is no genetic difference between the groups. Differences between the group means could no longer be partly explained by differences in genetic profiles. Thus, in this natural pairs design, the scores in the two groups would be paired up because they are identical twins.

The primary advantage of the natural pairs design is that it uses a natural characteristic of the participants to reduce sources of error. The primary limitation of this design is often the availability of participants. The researcher must locate suitable pairs of participants (such as identical twins) and must obtain consent from both participants.

Matched Pairs

In a natural pairs design, scores were paired for some natural reason. In a matched pairs design, scores are paired because the experimenter decides to match them on some variable. The rationale for the matched pairs design is the same as that for the natural pairs design—to reduce error variability by controlling extraneous variables.

Once again, let's return to our TV violence study. It is important for the researcher to consider possible matching variables prior to the study. As the researcher, you may decide that the gender and age of the child are critical variables that relate not only to the child's aggressive behavior, but also to how the TV program may affect them. You suspect that 5-year-old boys may be more aggressive, in general, than 3-year-old girls and may be more affected by the violence in a TV program. In the next chapter, we will see how these variables can be included as additional independent variables in the research design. But for now our goal will be to control them.

Instead of relying on simple random assignment to balance these variables (gender and age) across your groups, you begin by pairing participants in your sample. A 3-year-old girl is paired with another 3-year-old girl, a 5-year-old boy is paired with a 5-year-old boy, and so on. After all the pairs are created, you use random assignment to determine which of the participants in each pair will be in the experimental group and which one will be in the control group. Now the two groups are matched in terms of both age and gender. Differences between the group means can no longer be explained by differences in age or gender of the participants.

The primary advantage of the matched pairs design is to use experimental control to reduce one or more sources of error variability. One limitation of this design can be the availability of participants. At times, there may not be a suitable match for a participant. For this reason, the researcher should not try to match the groups on too many variables. The design can quickly become too difficult to manage. Usually one or two matching variables are sufficient. But remember, the matching variable(s) must be related to scores on the dependent variable. Otherwise, error variability will not be reduced.

Repeated Measures

With both the natural pairs and the matched pairs designs, our objective is to better equate the groups and to reduce random error due to individual differences. However, notice that we still have different participants in the different groups. Different participants will not only have different genetic backgrounds (unless they are identical twins), they will have very different sets of life experiences (including identical twins). These different life experiences shape a person and influence how he or she will behave in any given situation. Whenever you have different participants in the different experimental conditions, there will be some error variability due to individual differences. A solution is to use a repeated measures design, in which the same group of participants experiences all the conditions; that is, each research participant is tested under each treatment condition.

For our TV violence study, we would sample a group of children from day-care centers and then have them participate in both experimental conditions. On one day, the children would be observed after they had watched a TV program with violence. On another day, the same children would be observed

after they had watched a TV program without violence. To avoid confounding due to order effects, we would have to counterbalance the order of TV programs so that half the participants watch the violent program first and half watch the program without the violence first.

Advantages of Repeated Measures Designs. The beauty of this design is that it provides a means of controlling all of the extraneous variables associated with individual differences, including genetic background, socioeconomic status, age, gender, family structure, and type of parents. We have indicated that the greatest advantage of using a repeated measures design is the marked control over individual participant variation. Because each participant receives each treatment, participants with identical characteristics necessarily receive each of the different treatment conditions. Thus, any differences in performance should result only from the treatment conditions. In fact, however, this does not happen. Even though the same participant is used across treatments, the participant may change in some systematic fashion. The participant may be less observant or attentive from one treatment to the other, motivational levels may increase or decrease, fatigue or boredom may occur, or perceptions may change. Further, inevitable variations in the experimental setting, such as noise level or distractions, may affect performance. Therefore, because the participant and the environment may differ from treatment to treatment, there will still be some error variability, but far less than if an independent samples design had been used.

Another advantage of the repeated measures design relates to the population of available participants. If the availability of participants is low, then an independent samples design may not be possible. This predicament arises on occasion, especially when the population of interest is very small—for example, left-handed individuals with split-brain operations, identical twins separated at birth, or patients in therapy. The independent groups designs require k times as many participants as repeated measures designs (where k is the number of different treatments).

With fewer participants come greater efficiency and economy. In many cases, pretraining on a task may be needed one time only, after which a number of different treatments can be given. To illustrate, with four treatments and a task that requires a 10-minute pretraining period, a repeated measures design would save 30 minutes of training time per participant over an independent groups design. Thus, having only one training period may result in considerable savings in time, effort, and expenses. A similar savings can occur with instructions. In experiments involving different treatment conditions, the same instructions or similar instructions are commonly used. These instructions can be long and tedious. A repeated measures design can reduce the time devoted to instructions, particularly when instructions are the same across treatments.

A final important advantage is that a repeated measures design may be the most appropriate for the study of certain phenomena. It is the design of choice for studying learning and transfer of information,

or for assessing the effects of practice or repetition on performance. The independent variable is commonly the number of practice sessions given to individual participants. In this case, we are interested in the effects that earlier treatments have on later performance. In Chapter 8, we introduced the notion of carryover effects as a potential source of extraneous variability (confounder). However, they may be the phenomenon of interest to the researcher. Moreover, the concept of external validity enters into the choice of experimental design. The generalizability or representativeness of the research is related to the context in which it takes place, especially when the results of the research are to be used in applied settings. The setting in which the research takes place should be similar to the setting to which the experimenter wishes to generalize his or her results. It may be the case that the researcher is interested in situations where each individual receives a number of conditions or receives extensive practice. If so, then a repeated measures design would have greater external validity. On the other hand, if the researcher is interested in performance under conditions that minimize practice, an independent groups design is necessary.

A final example of a research project that lends itself to a repeated measures design is a **longitudinal research** study. As noted in the previous chapter, developmental psychologists are often interested in how the behaviors of individuals may change across portions of the life span or across the entire life span. We could ask the question "Do people's responses to violence on TV change as they develop from young children to school-age children to teenagers?" Instead of comparing preexisting age groups (cross-sectional research), such a study might involve repeated annual testing of the same children over a dozen years.

Methodological Issues With Repeated Measures Designs. A major methodological problem found with repeated measures designs is that they give rise to unwanted carryover effects. Any treatment other than the independent variable that changes the organism in such a way that it has a persistent effect on other treatments, we call carryover. We will distinguish three categories of carryover effects: (1) **transient effects**—short-term effects that dissipate with time; (2) permanent effects, most often due to learning; and (3) sensitization effects, resulting from experiencing all treatments. These carryover effects pose a problem for us when they are unwanted and their occurrence is confounded with the effects of treatment.

Short-term transient effects are often due to fatigue, boredom, or drugs. For example, let us assume that we are interested in evaluating the effects of Drugs A and B against a placebo condition using a psychomotor task involving coordination. We decide to use a repeated measures design in which each participant will receive each drug, including the placebo, in some random order. A tracking task is used in which the duration of contact with a moving target is recorded. Each participant is tested once each day, but under a different condition. Imagine that on Day 1, one drug was evaluated, and on Day 2, the

second drug was tested. What would happen if the effects of the first drug had not worn off? Performance on Day 2 would be a function of the second drug plus the persistent effects of the first drug. In short, the effects of one drug treatment would still be present when testing the effects of the other drug. Obviously, this is a case of blatant confounding, since we are not interested in the combined effects of the two drugs. This transient carryover effect can be easily corrected. Assuming that the changes in performance due to the drugs are not permanent, we could reduce this carryover by widely separating the treatments in time so that the previous drugs are out of the physiological system.

Another type of transient effect is that due to fatigue or boredom. Fatigue or boredom is especially likely to occur in nonchallenging studies requiring repetitive responding or in studies that take place over a long period of time. Therefore, when one treatment condition follows another, factors such as fatigue or boredom may contribute more to one condition than to the other. These factors would be mixed with our independent variable, thus making it impossible to evaluate. In short, we have confounding.

An example of the fatigue or boredom effect may help. Let us say that we are interested in evaluating the speed of responding to an auditory signal versus a visual signal. For one treatment condition, we use a five-second auditory signal and for the other, a five-second visual signal. Our dependent variable is speed of responding (pressing a key on a computer keyboard) to the two different signals. Participants first receive 100 trials of practice without any signal to assure that rapid responding will occur at the start of the experiment. All participants then receive the auditory signal first for 500 trials, followed immediately by the visual signal for another 500 trials. If we were to use the described procedure, we could not adequately evaluate the effects of signal modality. The possibility exists that the participants may experience fatigue, boredom, or both during the second 500 trials with the visual signal. If so, then there could be a systematic decrease in reaction time due to fatigue and/or boredom, thus resulting in our underestimating reaction time to a visual signal.

One way to avoid the problem of fatigue or boredom contributing more to one condition than the other is to use a **counterbalancing** procedure. Counterbalancing does not eliminate transient effects, but it allows us to distribute them evenly across the treatment conditions. It can be used easily with two treatments, less so with three, and only with difficulty with four or more treatments. Counterbalancing could be achieved in our reaction time experiment in several ways. The easiest way would be to have an equal number of participants receive the treatments in an A-B order as in a B-A order. A and B would represent either the visual or the auditory signal. It is important to note that the use of such a procedure assumes that the transient effects of fatigue or boredom when going from Treatment A to Treatment B are the same as the transient effects when going from Treatment B to Treatment A. If, in our example, the second treatment were more fatiguing or boring than the first, then our assumption would be in error. In this case, counterbalancing would not distribute the transient effects evenly for the two conditions. The

problem of equal treatment effects could be avoided and a repeated measures design still used by conducting the experiment over a two-day period. In this case, A and B would correspond to Days l and 2.

We previously noted that counterbalancing gets more difficult as the number of treatments increase. With two treatments, only two orders are possible: A-B and B-A. With three treatments, we have six possible orders: A-B-C, B-C-A, C-A-B, A-C-B, B-A-C, and C-B-A. However, with four treatments, we would have 24 orders, and with five treatments, we would be overwhelmed with 120 orders. When the number of treatments is greater than three, a random assignment procedure is far easier to use.

When repeated measures are taken on the same individual, we often see special kinds of permanent carryover effects. These are referred to as practice effects or learning effects. In many instances, practice effects are the independent variable of primary interest, but in other instances we try to avoid them. Practice effects can confound our research in ways that make our results uninterpretable. As we have noted, when our interest is in an independent variable other than practice, we must control practice effects so that they do not intrude on our results. In the preceding example where Drugs A and B were evaluated, we noted that the transient carryover effects of one drug on the other could be eliminated by widely spacing the time between tests. Knowing how long the drug remained active in the body would virtually assure us that we could eliminate transient effects. However, if for some unfortunate reason we did not randomize or counterbalance the presentation of drugs to each participant, a new problem would emerge. For example, if the effects of Drug A on the pursuit motor task were always tested first and the effects of Drug B were always tested second, then a marked practice effect (change in skill) could confound our results. Because trying to maintain contact with a moving target (pursuit motor task) is difficult, participants would initially do poorly on the task but would subsequently improve. Therefore, always practicing the task under Drug A first may lead to better performance under Drug B. However, the improvement may have little to do with the drug. The individual may simply now be more skilled because of practice. If our results came out the reverse, we could propose a reasonable alternative explanation—namely, that participants became more fatigued by the time of the second treatment. However, this argument could be weakened by lengthening the time interval between treatments.

Two things could be done to avoid practice effects. One would be to give sufficient practice on the pursuit motor task before giving any treatment condition. After improvement had stabilized or the limit of learning was attained, we could then introduce the treatments. This procedure would virtually assure that no increases in performance under the second treatment could occur as a result of practice. If our treatments were widely separated in time, we could also rule out fatigue factors. But the solution to the problem may create a new one if our interest is directed toward evaluating improvement in performance. If, because of our extended practice, participants are performing at their upper limits, further

improvement in performance as a result of our treatment may not be possible. This ceiling effect would obscure any enhancing effect on the pursuit motor task that the drugs might have. We would only be able to determine if they detracted from performance.

We could also deal with order effects by randomly assigning the order of treatments to each participant or counterbalancing them as described in the preceding section. When random assignment or counterbalancing is used, we assume that the effects of practice due to the order of presenting the treatments are the same for each treatment. If the carryover effects of practice are different, we then have confounded practice (order of presenting treatments) with the treatment effects. Whether this type of confounding has occurred can be determined by plotting performance across the different testing orders. Figure 12.1 illustrates the absence and presence of confounding due to order of presentation.

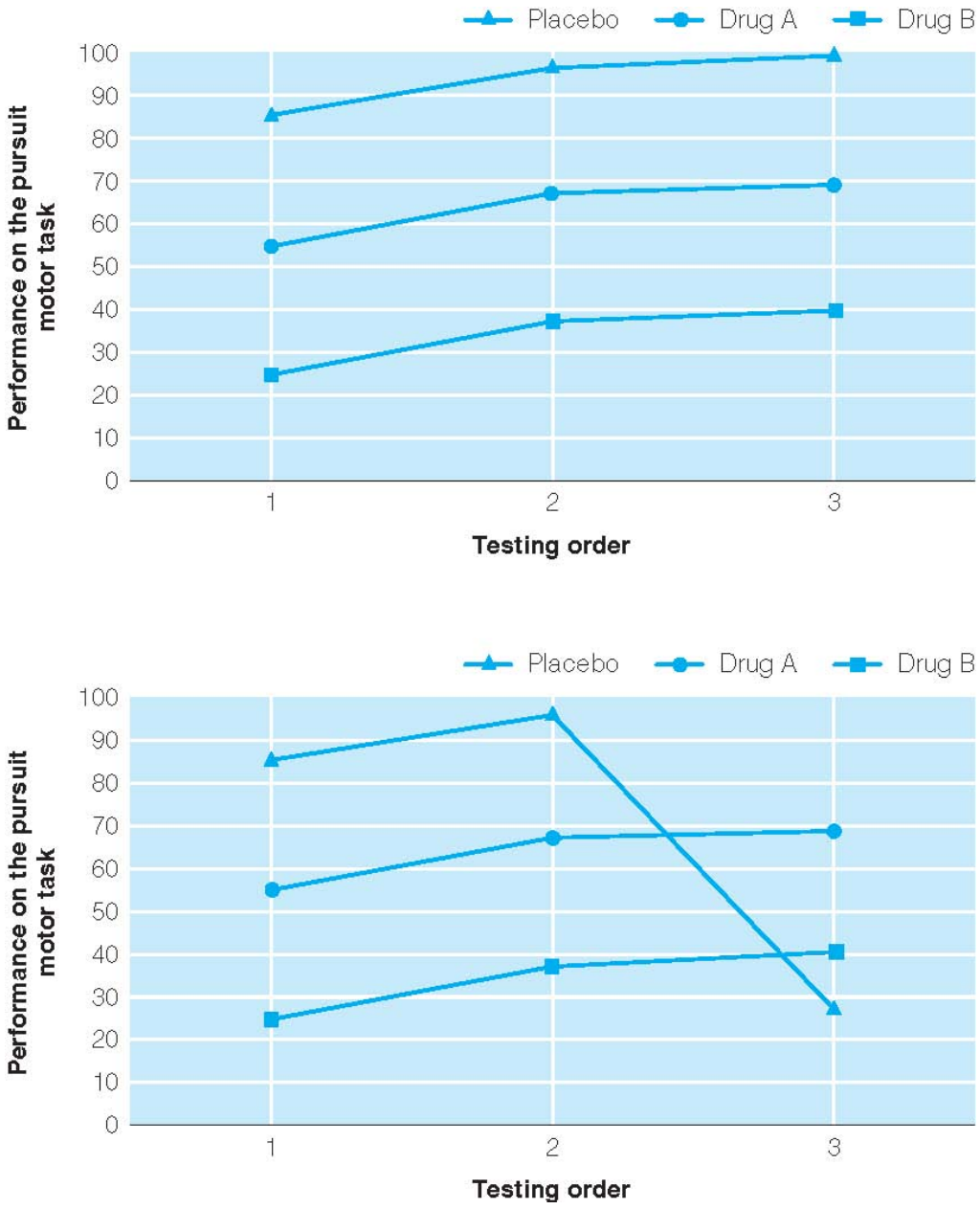

Figure 12.1 Graphic illustration of the absence or the presence of confounding due to order of presentation of treatment conditions.

If the results look like those at the top of the figure, then there is not a problem, because the practice effect is the same for each treatment whether it is given first, second, or third in the sequence. On the other hand, if the data look like those at the bottom of the figure, then we have confounded practice (that is, order) with the treatment effects. The bottom of the figure shows an interaction between the treatment conditions and the order of testing. What this means is that the practice effect is different for different treatments and the effect on performance that we observed is not a pure treatment effect. Clearly, the

order of presenting the conditions has some effect. Our performance measure reflects the effect of the treatments plus the practice due to the preceding treatment. Results such as this suggest that an independent samples design would be more appropriate.

A final category of carryover effects is referred to as sensitization. Experiencing the full range of treatments in an experiment may enhance participants' ability to distinguish differences in treatments and the extent of these differences. This may, in turn, allow the participants to contrast the various treatment conditions. Thus, their responses to a particular treatment may depend upon how they perceive that condition relative to the preceding one. Because participants are exposed to the entire range of stimuli when a repeated measures design is used, the context in which participants responds is very different from that of participants receiving only one treatment. Moreover, the demand characteristics are likely to differ from one design to another. After participants have received several treatments, they are more likely to form expectancies or hypotheses about the purpose of the experiment. If so, then these hypotheses may affect their performance over and above that of the treatments.

For example, sensitization effects may occur in a repeated measures design in which participants are asked to judge personality traits of persons pictured in photographs and the independent variable is the body size of the person in the photograph (overweight vs. not overweight). Although the order of the photographs could be randomized, it is likely that many participants would notice that the weight of the person is being manipulated and that the researchers are specifically studying how weight might affect personality judgments. This awareness of the relevant conditions might then affect their personality judgments on all remaining photographs (in a manner that might conceal a bias due to weight). Note that such a response by participants would not occur in an independent samples design.

Comparing Repeated Measures Designs With Independent Samples Designs. Because the context provided by exposure to all treatments is very different from the context provided by exposure to a single treatment, the participant's response to any given treatment may be, in part, a function of the research design. The most effective way to determine if different designs lead to different findings or behavioral laws is to compare experiments using repeated measures designs with those using independent samples designs. Determining the equivalence or nonequivalence of the two types of designs is important whenever different relationships are discovered and different designs are used to reveal them. This determination should also be made in other circumstances that go beyond methodological considerations. For example, it may be important for the construction of theories or for the application of findings to some practical problems.

Before we turn to the statistical analysis of correlated samples designs, there is one final statistical issue related to the comparison of independent samples designs and correlated samples designs. As we have noted in this chapter, the primary advantage of correlated samples designs is the reduction of

random error due to individual differences. This results in a larger value for either the *t* or *F* statistic and, therefore, a greater likelihood of detecting a significant treatment effect. However, the probability assigned to a particular *t* or *F* statistic also depends on the degrees of freedom associated with the analysis. The lower the degrees of freedom, the less likely we are to find a significant effect. For the independent samples design with two groups, the degrees of freedom are $(n_1 - 1) + (n_2 - 1)$ where *n* is equal to the sample size for each group. For the correlated samples design with two groups, the degrees of freedom are $(n - 1)$ where *n* is equal to the number of pairs of scores. For an experiment with 20 scores in each of two conditions, an independent samples design would have 38 degrees of freedom, whereas a correlated samples design would have 19 degrees of freedom. Thus, it is important for the reduction in random error associated with correlated samples to outweigh the reduction in degrees of freedom. This is generally the case for natural pairs and repeated measures designs because of the multitude of extraneous variables that are controlled. It is a more relevant consideration for matched pairs designs. If the matching variable does not substantially reduce the random error, the design is at a disadvantage.

Let's now take a look at the box "Thinking Critically About Everyday Information" and consider a repeated measures design that has some methodological problems.

Thinking Critically About Everyday Information: Effect of Frustration on Constructive Play in Children

Consider the following research report:

A researcher was interested in the effect of frustration on constructive play in children between the ages of 5 and 8 years. The hypothesis was that frustration would have an adverse effect on such play. Frustration was induced in the traditional way by thwarting or blocking performance of certain activities that children found pleasurable. A random sample of children in the proper age range was selected for study. The researcher then began the experiment with a 30-minute period during which the children played in the schoolyard in groups of ten. Groups of ten were used so that the experimenter could easily record both frequency and duration of constructive play. Then the children were brought into another condition, where frustration was induced. The children were then placed back into the original play situation, and frequency and duration of constructive play were again observed and recorded by the experimenter. Results of the study were unambiguous: Constructive play declined considerably following the frustration condition. Statistical tests revealed this outcome to be significant. The experimenter concluded that frustration was detrimental to constructive play, thus confirming the hypothesis.

Think about the following questions:

- What methodological issue is most problematic?
- What are some alternative explanations for the pattern of results?
- How would you improve the research design?

Comparing Two Groups

We have already mentioned how we might approach our TV violence study with a natural pairs, matched pairs, or repeated measures design. Because all three designs involve correlated samples, they can be

analyzed the same way. Let's return to our repeated measures design in which the same group of children watched both types of TV programs. Data that might be recorded from this study are shown in Table 12.2.

Table 12.2 Number of Aggressive Behaviors Recorded

TV Program With Violence (Experimental)					TV Program Without Violence (Control)				
8	9	0	10	4	10	2	2	2	0
7	3	2	8	8	1	3	3	5	7
6	12	8	4	4	2	0	2	6	8
5	6	5	5	9	11	2	4	6	4
9	15	12	3		7	7	9	5	
n = 24; M= 6.75; SD = 3.52					n = 24; M= 4.50; SD = 3.12				

Clearly, the mean number of aggressive behaviors when the children watched the TV program with violence (*Beast Wars*) is somewhat higher than when the children watched the TV program without violence (*Mister Rogers*). In addition to this variability between the groups, there is also variability within the groups such that scores were, on average, about 3 units from their respective means. The variability within the groups is due to random error, and the variability between the groups is due to any systematic error due to confounds plus any systematic variability due to the type of TV program. Again, the advantage of the repeated measures design is that individual differences do not contribute to the error between the groups of scores.

Variability Within Groups = Random Error (Extraneous Variables)
Variability Between Groups = Systematic Error (Confounds) + Systematic Variability (Effect of IV)

As with the independent samples design, the basic question is whether the difference between the two group means is due to error alone or due to error plus an effect of the independent variable (TV violence).

Comparing *t* Test to ANOVA

As with the independent samples design, the correlated design with two groups can be analyzed with either a **correlated samples *t* test** or a **one-way correlated samples ANOVA.** The correlated samples *t* test is sometimes referred to as a related samples *t* test or a paired samples *t* test. Likewise, the correlated samples ANOVA is often referred to as a repeated measures ANOVA. Recall from Chapter 10 that

parametric tests require assumptions of normality and homogeneity of variance. If there is reason to suspect that either of these assumptions is seriously violated in a correlated samples design, then a nonparametric test such as the Wilcoxon test is more appropriate. For the examples in this chapter, we will assume normality and homogeneity of variance.

Recall that *t* tests are restricted to the comparison of two groups, whereas ANOVAs can be used with two or more groups. In either case, the inferential statistic is based on a ratio of variability between groups to variability due to error.

$$\text{Inferential Statistic} = \frac{\text{Variability Between Groups}}{\text{Error Variability}}$$

Let's examine each one.

Correlated Samples t Test

The correlated samples *t* test uses the difference between the two group means as a measure of variability between groups and uses the standard error of the difference between means as a measure of error variability. The difference between the two group means is a straightforward calculation. The standard error of the difference between means tells you, on average, how different the two group means should be if the difference is due solely to error variability. If you examine the formulas in a statistics book, you will see that the standard error is based on the variability of the difference scores, where the difference scores are calculated for each pair of scores.

$$t = \frac{\text{Difference Between the Two Group Means}}{\text{Standard Error of the Difference Between Means (Error)}}$$

If the null hypothesis (H_0) is true—that is, there is no effect of the independent variable—then you would expect the difference between the two group means to be small and the *t*-statistic to be near 0. If, on the other hand, the null hypothesis is false, then you would expect the difference between the two group means to be large (in either a positive or a negative direction) relative to the standard error. The resulting *t*-statistic would have a value away from 0 (in either a positive or a negative direction). The larger the absolute value of the *t*-statistic, the lower is the probability that the difference between the group means is due solely to error variability. If the probability is low enough (what we refer to as the alpha level), then we reject the null hypothesis and accept the alternative hypothesis (H_1). We conclude that the independent variable had an effect.

For the data presented in this example, the output from a statistical analysis program would include the information in Table 12.3.

Table 12.3 Output From a Correlated Samples *t* Test

t	*df*	*p*
–2.52	23	0.019

This table shows that the *t*-statistic was –2.52, the degrees of freedom were 23, and the probability value was 0.019. Using an alpha level of .05, we decide to reject the null hypothesis and conclude that there was a significant effect of the independent variable on the dependent variable. Specifically, children who watched a TV program with violence showed significantly more aggressive behaviors than children who watched a TV program without violence, $t(23) = -2.52$, $p = 0.019$.

Correlated Samples ANOVA

As noted earlier, these same data could be analyzed with a correlated samples analysis of variance. As noted in the previous chapter, the logic of the ANOVA is very similar to that of the *t* test. Again we calculate a ratio of variability between the groups to error variability, referred to as the *F*-ratio. However, the numerator of the formula is not simply the difference between the group means. It is a measure of variability based on how different the group means are. Therefore, whereas the *t*-statistic can have negative values and has an expected value of 0, the *F*-ratio must be positive (because variability is always positive) and has an expected value of 1. Remember that expected values are based on the null hypothesis being true.

$$F\text{-ratio} = \frac{\text{Variability Between Groups (Mean Square}_{\text{between}})}{\text{Error Variability (Mean Square}_{\text{error}})}$$

Output from a computer program would include the information shown in Table 12.4.

Table 12.4 Output From a Correlated Samples ANOVA

Source of Variability	Degrees of Freedom (*df*)	Sum of Squares	Mean Squares	*F* Ratio	*F* Probability
Between groups	1	60.75	60.75	6.37	0.019
Within groups (error)	23	219.25	9.53		
Total	24	280.00			

As in the analysis using the *t* test, the probability of the *F*-ratio (0.019) is less than the alpha level (.05), so the decision would be to reject the null hypothesis and conclude that the independent variable had a significant effect on the dependent variable. Specifically, children who watched a TV program with violence showed significantly more aggressive behaviors than when the same children watched a TV program without violence, $F(1,23) = 6.37$, $p = 0.019$.

Comparing More Than Two Groups

Based on the above experiment, we concluded that children who watched a TV program with violence (*Best* Wars) showed significantly more aggressive behaviors than when they watched a TV program without violence (*Mister Rogers*). As with the independent samples design in the previous chapter, a control condition with no TV program would help us to determine which type of program is actually affecting aggressive behavior. The experimental procedures will be the same as previously described with the exception that there will be an additional condition in which the children watch no TV program in the 30 minutes prior to the observation period. This third condition will serve as something of a baseline with which we can compare the other two conditions. Let's add this third group to our hypothetical data (see Table 12.5).

Table 12.5 Number of Aggressive Behaviors Recorded

TV Program With Violence	TV Prograram With No Violence	No TV Program
8 9 0 10 4	10 2 2 2 0	7 7 10 4 3
7 3 2 8 8	1 3 3 5 7	12 6 6 5 8
6 12 8 4 4	2 0 2 6 8	2 11 12 3 3
5 6 5 5 9	11 2 4 6 4	5 9 9 11 7
9 15 12 3	7 7 9 5	10 4 5 4
M= 6.75; SD = 3.52	M= 4.50; SD = 3.12	M= 6.79; SD = 3.11

An inspection of the group means suggests that it may be the *Mister Rogers* program that reduced aggression in that group. You might also observe that the standard deviations are similar across the groups, thus supporting the homogeneity of variance assumption. We need to conduct analyses to tell us whether there are any significant differences, and if so, where they are. A *t* test is not an option because it is restricted to the comparison of two groups, and the use of multiple *t* tests is not an acceptable procedure because it inflates the Type I error rate. Therefore, a correlated samples ANOVA is the appropriate analysis. Output from a computer program would include the information in Table 12.6.

Table 12.6 Output From a Correlated Samples ANOVA

Source of Variability	Degrees of Freedom (*df*)	Sum of Squares	Mean Squares	*F* Ratio	*F* Probability
Between groups	2	82.53	41.26	3.87	0.028
Within groups (error)	46	490.14	10.66		
Total	48	572.67			

This output tells us that there is a significant difference among the three group means, $F(2,46) = 3.87$, $p = 0.028$. Because there are more than two groups, we cannot be sure where the differences are. To determine this, we conduct a post hoc specific comparison test (such as Tukey HSD or Sheffé). Output from a Tukey HSD shows that the means for Group 1 (TV Violence) and Group 3 (No TV) are significantly higher than the mean for Group 2 (No Violence). Therefore, we can now conclude that watching *Mister Rogers* for 30 minutes significantly reduced aggressive behavior when compared to watching *Beast Wars* or no TV. Notice the additional information that was provided by the multiple-group design.

Case Analysis

Let's consider a study in which the research participants experience both levels of an independent variable. An industrial/organizational psychologist is consulting with a large company that operates its factory 24 hours a day. The employees work on three rotating shifts: day shift (7 A.M.–3 P.M.), evening shift (3 P.M.–11 P.M.), and night shift (11 P.M.– 7 A.M.). Every month, employees rotate to a new shift. The research question is whether employee productivity is better with a clockwise rotation (day to evening to night to day) or with a counterclockwise rotation (day to night to evening to day). For the first six months, employees rotate clockwise, and for the second six months, employees rotate counterclockwise. The total number of production mistakes for each six-month period is recorded for 100 employees. Table 12.7 shows the descriptive statistics, and Table 12.8 shows the inferential statistics.

Table 12.7 Descriptive Statistics

Shift Work Rotation	*n*	*M*	*SD*
Clockwise	100	19.96	5.63
Counterclockwise	100	24.45	5.99

Table 12.8 Output From a Correlated Samples *t* Test

t	*df*	*p*
−5.20	99	<.01

Critical Thinking Questions

1. Based on the *t* test, is there a significant effect of the direction of rotation on employee mistakes?
2. Write a conclusion for the study that includes the direction of the effect.
3. Was this study a true experiment?
4. Can you conclude that the direction of rotation caused a change in worker productivity? Why or why not?
5. How could the study be improved so that the conclusion would be stronger?

General Summary

A correlated samples design is a true experiment characterized by assignment of participants to conditions in pairs or sets. The pairs or sets may be natural, matched, or repeated measures on the same participants. The design also includes manipulation of the independent variable. In conjunction with the use of control groups, this design permits cause–effect conclusions. Such conclusions are derived from the use of descriptive statistics and inferential statistics (*t* test, ANOVA).

The repeated measures design is quite common. Although this design has advantages, it also raises statistical and methodological issues. Advantages include a need for fewer participants and the ability to eliminate individual differences as a source of error between groups. Statistical issues involve assumptions of homogeneity of variance and covariance, wherein one assumes equal variability of scores in each of the treatment conditions and that participants maintain their relative standing in the different treatment conditions. Methodological issues include the effects of repeated testing—transient effects, permanent carryover effects, and sensitization. Counterbalancing techniques can be used to address the methodological issues.

Now that we have a fundamental understanding of experimental designs with one independent variable, the next chapter will explore designs with multiple independent variables.

Detailed Summary

1. Correlated samples designs do not use random assignment of participants to conditions. Rather, scores in the groups are paired up (assuming two groups) because they are natural pairs of participants, matched pairs of participants, or repeated measures from the same participants.
2. Correlated samples designs involve strategies to equate the comparison groups on variables other than the independent variable so that any differences between the comparison groups can be attributed to the manipulation of the independent variable.
3. Natural pairs designs involve the use of a natural variable (such as twins, siblings, or married couples) to equate the comparison groups. Good examples are identical twin studies in which one twin is randomly assigned to one of the treatment conditions and the other twin to the other treatment condition. These natural pairs equate the groups in terms of the genetic profiles of the participants.
4. The primary advantage of the natural pairs design is to use a natural characteristic of the participants to reduce one or more sources of error between the groups. The primary limitation of this design is often the availability of participants.
5. Matched pairs designs involve the use of an experimenter-chosen variable to equate the comparison groups. After pairs are established, one participant from each pair is randomly assigned to one treatment condition and the other participant to the other treatment condition.
6. The primary advantage of the matched pairs design is to use experimental control to reduce one or more sources of error between the groups. One limitation of this design can be the availability of participants. At times, there may not be a suitable match for a participant.
7. Repeated measures designs involve the repeated testing of the same participants such that each participant experiences all treatment conditions. This procedure eliminates error variability due to individual differences between the groups. Other advantages include efficiency, economy, and the ability to study phenomena that lend themselves to repeated testing (such as learning and practice).
8. Methodological concerns with repeated measures designs focus on three categories of carryover effects: (1) transient effects—short-term effects that dissipate with time; (2) permanent effects, most often due to learning; and (3) sensitization effects, resulting from experiencing all treatments. These carryover effects pose a problem when their occurrence is confounded with the effects of treatment.
9. All three types of correlated samples designs are analyzed in the same way. A two-group study can be analyzed with either a correlated samples *t* test or a correlated samples ANOVA. A multiple-group study must be analyzed with a correlated samples ANOVA.

Key Terms

carryover effects

correlated samples t-test

counterbalancing

longitudinal research

one-way correlated samples ANOVA

transient effects

Review Questions / Exercises

1. Summarize the essential characteristics of a one-way correlated samples research design.
2. Briefly describe a matched pairs experiment for which a correlated samples *t* test would be the appropriate inferential statistic. The experiment should test which is more effective in treating depression, a behavior modification program or a cognitive therapy program.
3. Briefly describe a repeated measures experiment for which a repeated measures samples ANOVA would be the appropriate inferential statistic. The experiment should test the effect of 4-hour food deprivation, 8-hour food deprivation, and 12-hour food deprivation on how fast a rat will run through a maze to obtain food.
4. In your own words, describe the methodological and statistical advantages of the repeated measures design. Also describe the statistical disadvantage and the potential methodological disadvantages.
5. Describe a repeated measures experiment related to human memory for which counterbalancing would be an essential methodological tool.