

One Variable With Two Related Groups

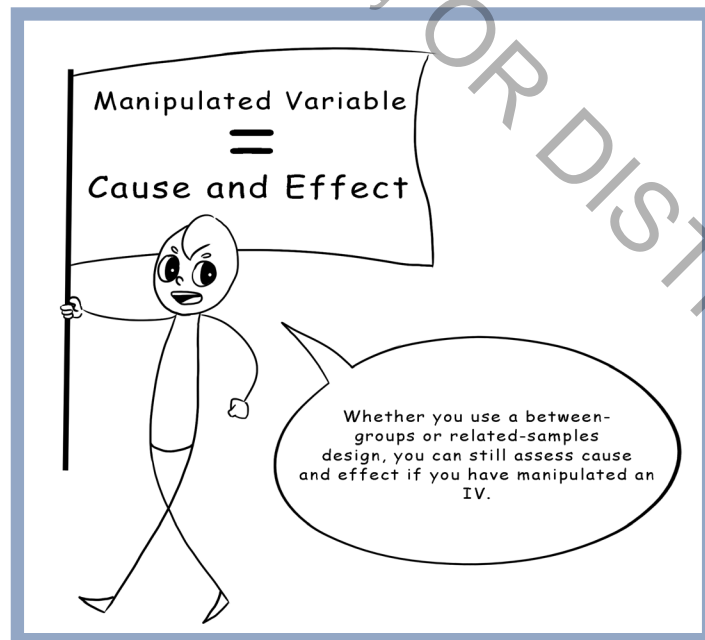
In chapters 10 and 11, we discussed one independent variable (IV) or quasi-IV with two or more levels. With both designs, each group contained different people in levels of the IV. As an alternative, some researchers design studies in which they compare people related in some way prior to entering the study, such as siblings. In this chapter, we will discuss examples of **related-samples designs**. Related-samples designs are also called **dependent-samples, paired-samples, and matched-pairs** designs. We will focus on an independent variable with only two levels, so you might guess correctly that we will use a *t*-test for analysis. Specifically, we will analyze data using a **related-samples *t*-test**.

Related-Samples (Dependent-Samples, Paired-Samples, Matched-Pairs) Design

A related-samples design is defined by testing the same, similar, or matched participants across IV or quasi-IV levels.

Related-Samples *t*-test

The related-samples *t*-test is the statistical test used when a design has the same, similar, or matched participants in two levels of an IV or quasi-IV with one interval or ratio DV.



Research designs that allow similar people to be measured across all levels of the IV or quasi-IV are powerful. But why? If each pair of participants is similar before arriving at your study, individual differences are minimized. Think about it. Suppose we want to know if giving a speech to an audience causes a larger increase in heart rate than giving a speech to an empty room. If we study pairs of siblings, we can put one sibling in each IV condition. At the end of the study, any differences in heart rate likely could be attributed to the IV conditions rather than individual differences because siblings should have similar heart rates. In comparison, when completely unrelated people exist in the two groups, differences in heart rate could be caused by the IV levels or a lot of potential individual differences such as physical fitness. Sure, you could argue that even sibling pairs can differ on heart rate, but it is reasonable to assume that siblings would not differ as much as two completely random people pulled from a population.

TESTING THE SAME PEOPLE TWICE

Testing pairs of people, with one in each IV level, certainly minimizes individual differences across conditions. But researchers can use an even stronger research design. The same participants can experience *both* IV levels. After all, the person most similar to you is you!

Let us practice using an example with two conditions. Recent evidence suggests that children are more likely to believe information given by an attractive woman than an unattractive woman (Bascandziev & Harris, 2014). The researchers showed preschool children pictures of novel objects, then had two women who varied in attractiveness say different names for the objects. The children picked the woman more likely to be correct, creating a simple frequency DV to be analyzed with a nonparametric statistic. We might design a follow-up study with an interval DV to allow analysis using a parametric statistic, which we know to be more powerful. We could assess children's confidence in each woman's answer on a Likert-type scale from 1 to 7, with higher numbers indicating more confidence.

How sure are you that the woman knows the right name for this object?

I'm sure she does not know							I'm sure she does know
1	2	3	4	5	6	7	

Consider the research question: *Do preschool children have more confidence in an attractive woman than an unattractive woman?* Stated as a directional research hypothesis, this question becomes, *Preschool children have more confidence in an attractive woman than an unattractive woman.* The variables must be operationalized, with pictures of an attractive woman and an unattractive woman offering two IV levels.

You might be thinking that beauty is subjective, and you would be right. To make sure we choose pictures of women who most children agree are attractive or unattractive, we could

show a small group of preschool children five pictures of women and ask them to order the pictures from least attractive (“ugliest”) to most attractive (“prettiest”). The children who rank the pictures are not part of the sample addressing our research question, but they do provide valuable preliminary information when designing the study. We might analyze these data by giving points for each ranking (e.g., 1 point for least attractive and 5 points for most attractive), calculating a mean score for each woman, and choosing pictures with the lowest and highest attractiveness means. The two pictures could then be used in the study.

We have already decided to test the same children twice, asking them to rate confidence in the unattractive and attractive women. Because we can manipulate participants’ actions by asking them to look at either the attractive or unattractive woman, the two conditions represent a true IV, and we can examine cause and effect. In prior chapters, random assignment to IV levels helped remove potential confounds, but testing the same people in both conditions means we cannot use random assignment here. Instead, we might ask children to view an unusual object, tell them what the *attractive* woman called the object, and have them circle a number on our confidence scale to indicate how confident they are that the woman correctly named the object. Next, the same children could be asked to view a second unusual object, followed by a label provided by an *unattractive* woman. Children again would rate their confidence using the 7-point scale. We would use a statistic to assess a potential difference between the two conditions.

But wait. Do you see any potential problems with the research design? We have outlined a method in which children always view the attractive woman first and the unattractive woman second. Differences in confidence across the two conditions might be explained by confidence in the women, but other reasonable explanations are possible. For example, children might be bored by the time they view the second woman’s picture, causing them to have less confidence in the unattractive woman. Or they might distrust the unattractive woman because they are comparing her with the attractive woman they saw first. Many explanations are possible when the same people are tested more than once in the same order of conditions. In fact, this is such an important concern that researchers have created labels for potential problems.

Order Effects (Carryover Effects)

Order effects occur when participants are influenced during the study by levels of the IV they already experienced. With a two-condition design, an order effect means the first condition influenced responses on the second condition.

PROBLEMS WITH TESTING THE SAME PEOPLE TWICE

When the same participants are tested more than once, several problems can be associated with changes in the participant across time. The first manipulation and assessment can change participants’ responses to the second manipulation. In general, these problems are called **order effects** because outcomes are impacted by the order of conditions. Order effects are also called **carryover effects** because the effect of one condition carries over to affect the next condition. Order effects occur when participants are influenced during the study by levels of the IV they already experienced. With a two-condition design, an order effect means the first condition influenced responses on the second condition. We do

not want this problem. Returning to our example, if children in the study always view the attractive woman first, their mood might improve across the duration of the study. When they experience the second condition, viewing the unattractive woman, they might rate their confidence in her as high because they are in a particularly good mood. We will not get a clear measure of confidence in the unattractive woman because mood improved in the first condition and carried over to the second condition.

Order effects may be based solely on which condition comes first, but the term also encompasses practice effects and fatigue effects. Alternatively, anything that occurs *between* the first and second manipulation might impact the final DV measure. In other words, people change across time for many reasons that are not tied to our study.

Practice Effect

In some studies, completing a DV the first time allows practice and may improve performance when the DV is completed a second time, defining a **practice effect**. Although practice effects likely would not be a problem in the current study, researchers consider the possibility when DV performance can improve with practice (e.g., quiz performance). Practice effects are a specific type of order effect.

Fatigue Effect

A second type of order effect is fatigue. When participants get tired or bored across a repeated assessment, their performance on the DV may suffer, causing a **fatigue effect**. After the first condition is experienced, and the DV is completed the first time, motivation and energy may decrease. In the current study, children may experience fatigue after examining an ambiguous object and deciding how confident they are in a woman's identification of the object. By the time they see a second object and rate confidence in the second woman, children may not give as much thought to their confidence ratings.

History Effect

Beyond order effects, testing the same people twice allows the possibility of an event between IV levels. A **history effect** occurs when participants change due to anything that occurs across the study. For example, suppose children viewed the attractive woman, and then the fire alarm at school forced everyone to leave the building for 30 minutes. When the children returned, you could continue the study by showing the unattractive woman. You can imagine that children might be affected by the excitement of a fire alarm and standing on the front lawn. A historical event between the two levels can alter outcomes on the second level.

Of course, history is most likely to be a problem when a large event, like a devastating hurricane, occurs between IV levels. Or history effects would be a concern if a long period of time elapses between IV levels, as may be the case in a study that assesses attitudes toward a teacher at the beginning of the term and at the end of the term. Researchers must

Practice Effect.

In some studies, completing a DV the first time allows practice and may improve performance when the DV is completed a second time, characterizing the practice effect.

Fatigue Effect.

A specific type of order effect is fatigue. When participants get tired or bored across a repeated assessment, their performance on the DV may suffer, revealing a fatigue effect. After the first condition is experienced, and the DV is completed the first time, motivation and energy may decrease.

History Effect.

A history effect occurs when participants change due to anything that occurs across the study. A historical event between exposure to the two IV levels can alter outcomes on the second level.

recognize the potential for history to compromise internal validity, and we might feel less confident that the IV caused changes in the DV.

Maturation

A final potential problem with repeatedly measuring the same participants is maturation. **Maturation** refers to the fact that people age and change over time. Suppose in our example we wanted to examine confidence over time by testing the same children both in preschool and in sixth grade. We would not want to give the attractive condition in preschool and the unattractive in sixth grade, for example, because getting older might change the way children view attractiveness.

Maturation.

Maturation refers to the fact that people age and change across IV or quasi-IV levels when the levels occur far apart in time.

Solving Order Problems by Counterbalancing

What is the solution to order effects, history effects, and maturation? Whenever possible, counterbalance the order of conditions. **Counterbalancing** usually is accomplished by randomly assigning people to *order* of IV levels. In our attractiveness example, approximately half of the children should view the attractive woman first and rate their confidence in her answer before moving on to the unattractive woman. The remaining participants should view the unattractive woman first. Counterbalancing the order of IV levels equally distributes any potential order effects. You can be more confident that changes in the DV are due to IV levels rather than which level came first.

Counterbalancing.

Counterbalancing is accomplished by randomly assigning people to order of IV levels. This technique defends against order effects, history effects, and maturation.

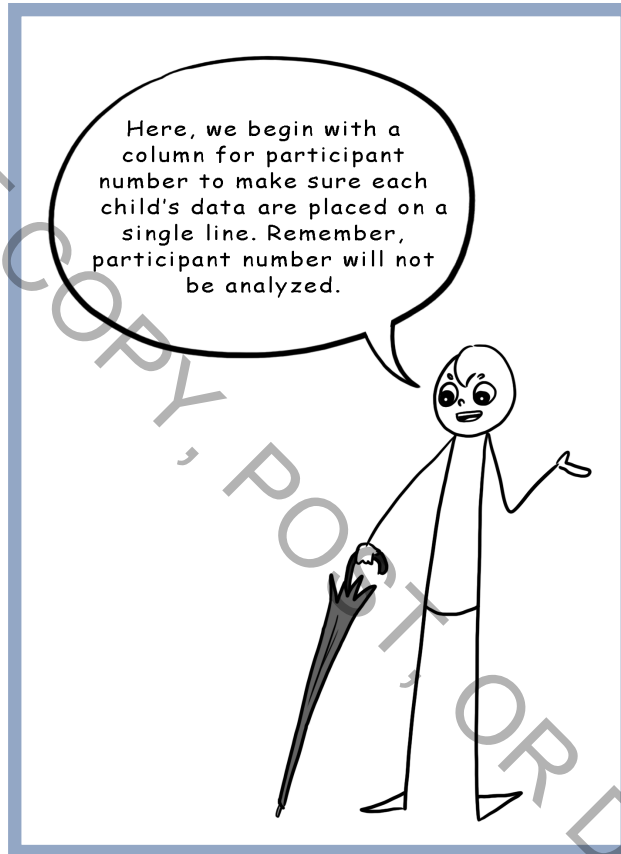
Avoiding Confounds

Remember from Chapter 3 that a confound is a variable that changes exactly along with the IV levels. In our example, we have improved our design by counterbalancing the order of conditions, but we still have a potential confound. If the picture of the attractive woman is always shown with one specific ambiguous item, and the unattractive woman is always shown with a second ambiguous item, we will not know if the woman's picture or the item altered confidence ratings. What if the item identified by the attractive woman is simpler, clearer, or somehow familiar to the children? The item could instill confidence. To avoid this potential confounding variable, we could randomly assign which object is paired with which woman. As an alternative, we could systematically pair a specific object with the attractive woman for half of the participants and pair it with the unattractive woman for the other half. As you can see, designing a study requires careful consideration of many details.

RESEARCH DESIGN: ONE IV WITH TWO RELATED GROUPS

Let us examine fictional data from our study of children's confidence in an attractive versus an unattractive woman. We have a DV that ranges from 1 (*I'm sure she does not know*) to 7 (*I'm sure she does know*). Keeping in mind that we tested the same children twice, we must

enter confidence values for each child on a separate row. We also have to be careful to put values in the correct column given that the order of conditions varied across participants.



Participant	Attractive	Unattractive
1	5	3
2	6	6
3	7	3
4	4	2
5	6	6
6	5	3

(Continued)

(Continued)

Participant	Attractive	Unattractive
7	6	5
8	7	5
9	6	4
10	7	5
11	4	2

SPSS: Related-Samples *t*-Test (Experimental Design)

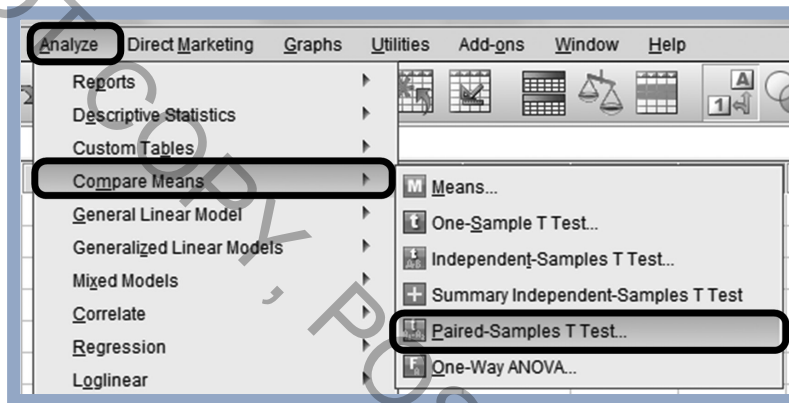
To enter these data into SPSS, go to Variable View and label column headings. When we had different people in our groups, we had one column for the IV (or quasi-IV) and one for the DV. With repeated measures, we enter data differently. Now we will give each level of the IV its own column. Label a column for Participant, a second column for Attractive, and a third column heading for Unattractive. We could enter value labels corresponding to our anchors, but labels are not needed to understand the data. Higher numbers reflect more confidence.

Name	Type	Width	Decimals	Label	Values
Participant	Numeric	8	2		None
Attractive	Numeric	8	2		None
Unattractive	Numeric	8	2		None

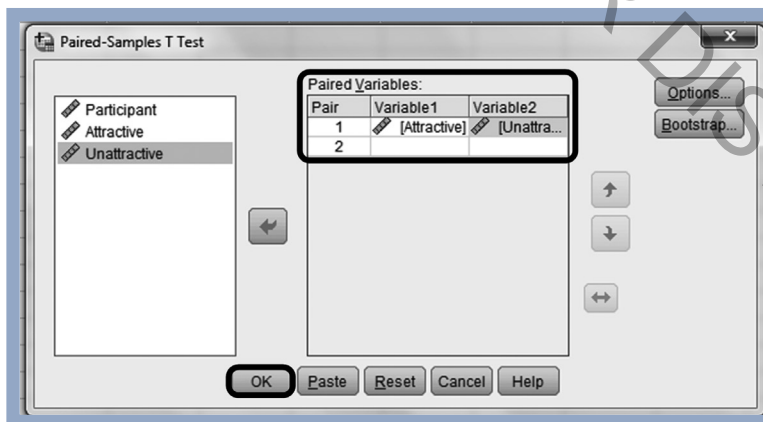
Go to Data View and enter the data exactly as they are shown in the original data table.

Participant	Attractive	Unattractive
1.00	5.00	3.00
2.00	6.00	6.00
3.00	7.00	3.00
4.00	4.00	2.00
5.00	6.00	6.00
6.00	5.00	3.00
7.00	6.00	5.00
8.00	7.00	5.00
9.00	6.00	4.00
10.00	7.00	5.00
11.00	4.00	2.00

Regardless of whether participants see the attractive or unattractive picture first, confidence rating with the attractive picture is listed first for data analysis, and the confidence rating for the unattractive picture is typed second. You could have entered the data in either order as long as you are careful to put the correct DVs in each column. Click Analyze, Compare Means, Paired-Samples T Test. SPSS refers to this research design as “paired” because the two values for each participant appear on the same row. You learned in Chapter 10 that although the SPSS term is “T Test,” researchers generally call the statistic a *t*-test.



In the box that opens, move Attractive to the right using the arrow. Attractive will appear in the Paired Variables box under Variable 1 beside Pair 1. Next move Unattractive to the right. It will appear beside Attractive, under Variable 2. Clicking OK allows SPSS to compare DV values across the two conditions.



Part of the SPSS output is shown in the following screenshot. From the first table visible, you will need descriptive statistics for the APA-style results section as well as for graphing, if you choose to include a figure.

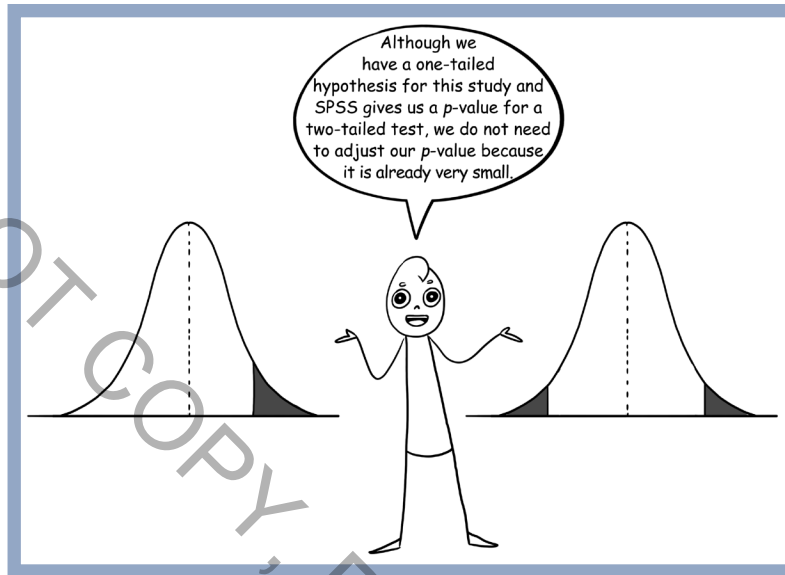
T-Test					
Paired Samples Statistics					
		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Attractive	5.7273	11	1.10371	.33278
	Unattractive	4.0000	11	1.48324	.44721

Further down in the output you will see the t -test outcome, degrees of freedom (df), and p -value. The df value is calculated using number of pairs of scores minus 1 ($11 - 1 = 10$).

Paired Samples Test					
Paired Differences					
Std. Error Mean	95% Confidence Interval of the Difference		t	df	Sig. (2-tailed)
	Lower	Upper			
.33278	.98579	2.46878	5.190	10	.000

The table labeled Paired Samples Test shows whether or not the two groups significantly differed. At the far right side of the table, look at the Sig value to see if it is $p \leq .05$. With a significance value of $p < .001$, we can say that the confidence ratings in the two conditions were meaningfully different. We can reject the null hypothesis: *Preschool children had similar confidence in an attractive woman and an unattractive woman.* Instead we found evidence in support of the research hypothesis: *Preschool children had more confidence in an attractive woman than an unattractive woman.*

We revealed a significant effect using a nondirectional test, but because we expected more confidence with the attractive woman than the unattractive woman, we actually needed a directional test. Of course, a directional test has more power and merely requires dividing the p -value by 2. In this case, $.000/2$ is silly, so we stick with reporting $p < .001$. If we return to the descriptive statistics, the means show that children rated more confidence in the attractive woman than in the unattractive woman.



Confidence Intervals

In the same output table showing the t -test result, notice the 95% confidence interval for the difference between the two groups (below). Values in the population likely range from a mean difference of 0.99 to 2.47, as shown by the confidence limits for this example.

Paired Samples Test		
Paired Differences		
Std. Error Mean	95% Confidence Interval of the Difference	
	Lower	Upper
.33278	.98579	2.46876

Effect Size: Cohen's d

With a significant effect, APA style requires effect size. As you learned in Chapter 10, SPSS will not include t -test effect sizes. We again turn to Cohen's d to communicate the size of an effect if the outcome is significant. The formula for Cohen's d with a paired-samples t -test is below.

$$d = \frac{M_{group1} - M_{group2}}{SD}$$

The numerator is easily located on the SPSS output. Look at the table for Paired Samples Test. Under Mean you will find the mean difference (Mean), which is one group mean subtracted from the other. Standard deviation (Std. Deviation) is immediately to the right in the table, as shown here.

Paired Samples Test					
		Paired Differences			
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval Lower
Pair 1	Attractive - Unattractive	1.72727	1.10371	.33278	.98579

Complete the formula as follows.

$$d = \frac{1.72727}{1.10371} = 1.56$$

Divide the numerator by the denominator to obtain 1.56. Recall from Chapter 10 that a Cohen's *d* of approximately .20 is considered a weak effect size, .50 is a moderate effect size, and .80 is a large effect size (Cohen, 1988).

APA Style for the Related-Samples *t*-test: Experimental Design

Although the wording in the results section below does not explicitly say “cause” or “effect,” participants were manipulated, allowing a discussion of cause and effect. This example represents a true experiment. The IV was which picture the students viewed.

Method

Participants

Children ($N = 11$) in a 3-year-old preschool classroom in Atlanta, Georgia, participated in this study. Age averaged 3.71 years ($SD = 0.27$), and ethnicities included 5 Black, 4 White, and 2 undisclosed ethnicities. All participants received ethical treatment, and the IRB approved the method.

Materials

Choice of stimuli. Prior to data collection for the study, 7 children from a 3-year-old preschool classroom viewed five pictures of women and ordered them from least attractive (“ugliest”) to most attractive (“prettiest”). Pictures received points according to their ranking (e.g., 1 point for least attractive, 5 points for most attractive), and we calculated

mean scores for each picture. The pictures with the lowest ($M = 1.49$, $SD = 0.78$) and highest ($M = 4.50$, $SD = 1.05$) scores served as stimuli for the study.

Confidence scale. To assess children's confidence in the woman's answer, we asked them, "Do you think this woman knows the right name for this object?" Children indicated confidence in the woman's object name on a scale from 1 (*I'm sure she does not know*) to 7 (*I'm sure she does know*).

Procedure

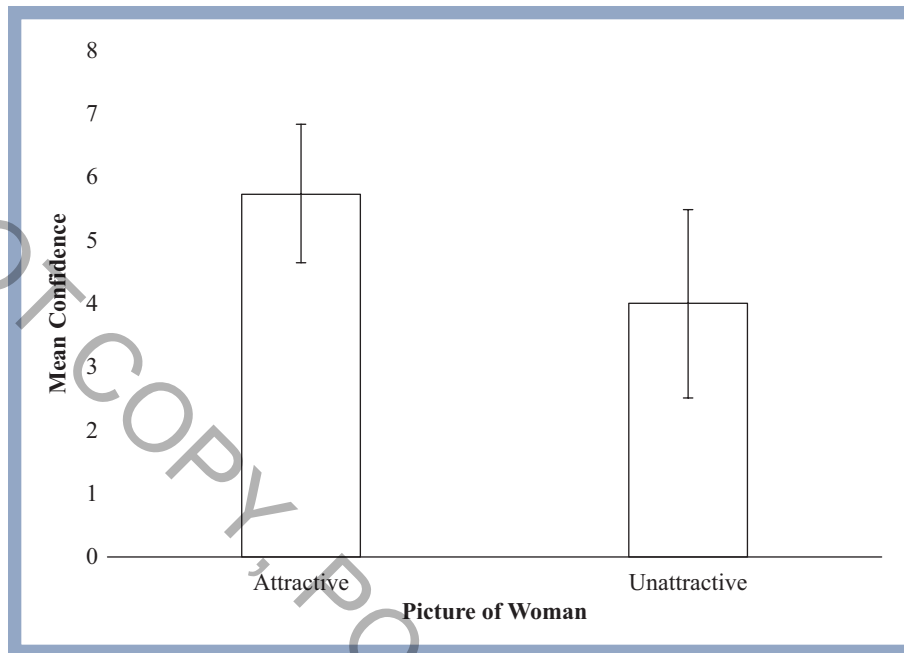
We approached parents at the beginning of the school day and asked them to sign a consent form allowing their child to participate in the study. At that time, parents also completed a form indicating demographics for their child. Throughout the day, researchers tested children individually in a quiet room. After asking if the children assented to the study, researchers showed them a picture of a novel object alongside a picture of a woman and indicated a name of the object given by the woman. Children rated their confidence in the object's label. Children next saw a picture of the second woman and answered the same question about a different object's label. We counterbalanced the order of pictured women, providing a random order for each participant. We also randomized which object and object name appeared with each woman's picture. Children received a sticker for participation.

Results

We analyzed these data using a one-tailed, paired-samples t -test. Children rated their confidence in a woman's object identification differently based on her level of attractiveness, $t(10) = 5.19$, $p < .001$, 95% CI [0.99, 2.47], $d = 1.56$. When children believed information came from an attractive woman, they rated their confidence in the information higher ($M = 5.73$, $SD = 1.10$, $n = 11$) than when they thought the information came from an unattractive woman ($M = 4.00$, $SD = 1.48$, $n = 11$).

You may have noticed that in a paired-samples t -test, the same number of participants will be in both conditions because each participant is tested twice. You will always have the same number of people in both levels, and therefore it is not really necessary to write $n = 11$ for each condition in the APA-style results section. We could have indicated the sample size of 11 once in the results section.

As we have noted before, you may want to include a figure in your manuscript. If so, refer to the figure in the APA-style results section. In this example, the IV is attractiveness based on two categories: a picture of an attractive woman and a picture of an unattractive woman. As you know, categorical data represent a nominal variable. A nominal IV is graphed using a bar graph as shown here.



Power

We were fortunate to reveal a significant effect in this study, especially with a small sample size. Power analysis would have revealed the need for 34 participants. In fact, for all paired-samples t -tests, you will need 34 participants for adequate power to detect a medium effect size at $p \leq .05$, as depicted in the table below. In our example, we used fewer participants for simplicity, but you should strive for enough participants to enhance power.

	Small Effect Size	Medium Effect Size	Large Effect Size
Related-samples t -test	200	34	16

Note: Numbers in the table represent **total** sample size. Double the number if you have different people in the IV or quasi-IV levels (e.g., siblings).

RESEARCH DESIGN: ONE QUASI-IV WITH TWO RELATED GROUPS

In the prior example, we examined a related-samples design in an experimental study, but this approach can be used just as easily in a correlational design with no manipulation. Suppose we wanted to know how high-school women feel about their athletic ability currently as compared to how they felt in their elementary-school years. We found an article claiming that among college athletes, comments about gender stereotypes in sports led women to underperform on an athletic task as compared to when a gender-stereotype

statement was not made (Hively & El-Alavli, 2014). We might wonder if, over time, girls start to internalize gender stereotypes about sports. Asking high-school girls to think back to their feelings of competence in elementary school might sacrifice accuracy. It would be better to collect information from elementary-school girls, and then have them rate competence again in high school. But waiting is not always realistic. Instead researchers often rely on asking people to recall information.

Based on the research by Hively and El-Alavli (2014), we can devise a research question: *Do high-school girls report less athletic ability for their current age than when they were in elementary school?* The research hypothesis offers a statement: *High-school girls report less athletic ability for their current age than when they were in elementary school.* Is this design an experiment? No. We are merely asking girls to report information and are not manipulating them in any way. This design uses a quasi-IV: elementary-school and high-school time periods. We will not learn cause and effect, but we will learn if time period relates to feelings of athletic competence. Thus, our research design is correlational. We collect two pieces of data from the same people, and analysis will require a paired-samples *t*-test.

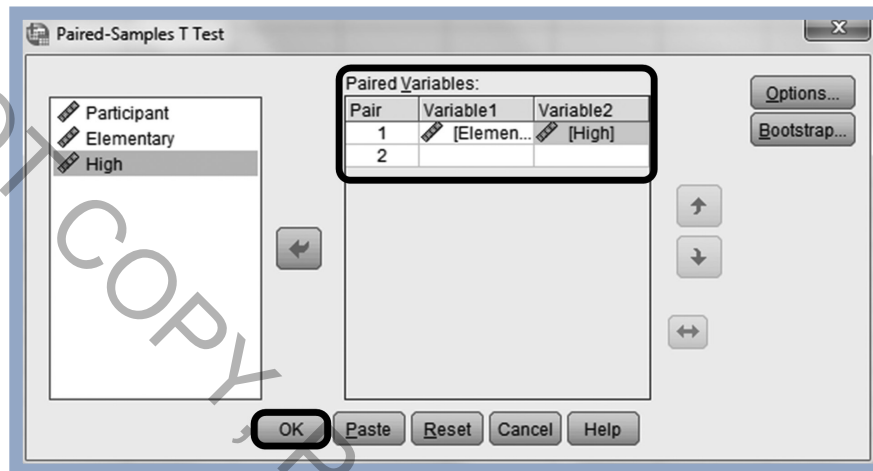
Now that we know our research question, design, and planned statistical analysis, we can examine hypothetical data. Below is a table of athletic ability ratings using a scale from 1 to 100, with higher numbers indicating more perceived ability for their age. We have included Participant Number as a column to help keep the data organized by row. Although we would have needed 34 participants for acceptable power, we will use a small data set for a concise example.

Participant Number	Elementary School	High School
1	80	65
2	71	58
3	90	74
4	50	60
5	88	69
6	45	45
7	36	22
8	95	82
9	50	55

SPSS: Related-Samples *t*-test (Correlational Design)

In Variable View, enter Participant, Elementary, and High as the three column headings. Under Data View, enter the data as shown in the above data set (review the prior *t*-test

example if needed). Under Data View, click Analyze, Compare Means, Paired-Samples T Test to analyze these data. In the box that opens, click Elementary and High over to the right side. Do not analyze Participant in the t -test.



Partial SPSS output is shown below. We have circled the relevant information to compare the two groups and report descriptive statistics.

Paired Samples Statistics					
		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Elementary	67.2222	9	22.25296	7.41765
	High	58.8889	9	17.58156	5.86052

The t -test table shows whether or not the two groups significantly differed. Notice also the arrow pointing to the 95% CI.

Paired Samples Test					
Paired Differences			t	df	Sig. (2-tailed)
Std. Error Mean	95% Confidence Interval of the Difference				
	Lower	Upper			
3.48807	.28982	16.37685	2.389	8	.044

First we examine the Sig value to see if it is no higher than .05. With a .044 p -value, we can say the two groups differed in a meaningful way. We can reject the null hypothesis: *High-school girls reported the same athletic ability for their current age as when they were in elementary school.* Our result provides evidence for the research hypothesis: *High-school girls reported less athletic ability for their current age than when they were in elementary school.*

As you might have realized, we need to divide the p -value by 2 because we expected high-school girls to perceive lower athletic ability currently than when they were in elementary school. Thus, our directional p -value is .022. Effect size is needed, requiring us to divide the mean difference of 8.33 by the standard deviation of 10.46 to obtain $d = .80$, a large effect size.

APA Style for the Related-Samples t -test: Correlational Design

In an APA-style results section, we will report the t -value, df , p -value, confidence interval, and Cohen's d . We also need to share with the reader which age group rated their athletic ability higher. Look at descriptive statistics in the first box of the SPSS output to locate means and standard deviations for the two age groups.

Method

Participants

Nine female high-school juniors participated in this study. Ages ranged from 15.07 to 16.98 ($M = 16.50$, $SD = 0.78$), and ethnicities included 7 White and 2 Black individuals.

Procedure

Guidance counselors invited students who worked in the school office to participate. Parents provided informed consent prior to data collection. In addition, students provided assent to participate when they visited the guidance office. Students rated their current perceived athletic ability relative to their peers on a scale from 1 to 100, with higher numbers indicating better perceived ability. Using the same scale, they reflected on their athletic ability relative to their peers when they attended the fifth grade.

Results

We analyzed these data using a directional paired-samples t -test. Age level related to confidence in athletic abilities among women, $t(8) = 2.39$, $p = .022$, 95% CI [0.29, 16.38], $d = .80$. Women in high school perceived their athletic abilities to be lower ($M = 58.89$, $SD = 17.58$, $n = 9$) than when they attended elementary school ($M = 67.22$, $SD = 22.25$, $n = 9$).

RESEARCH DESIGN: TESTING DIFFERENT PEOPLE (MATCHED PAIRS)

In the examples above, we tested the same people twice. We can also use the paired-samples *t*-test to assess similar people. One way to pair similar people is to test sibling pairs, as mentioned at the beginning of this chapter. Another option involves matching participants based on some characteristic related to what we are studying. Researchers call the latter approach a matched-pairs design.

Matching Participants

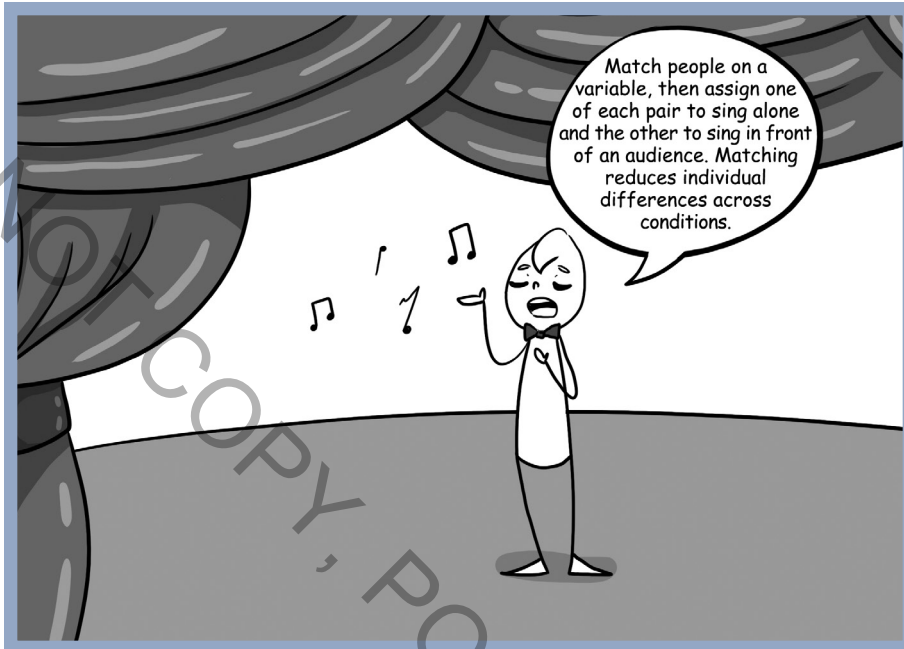
You saw earlier in this chapter that assessing the same people in all conditions has its own set of problems, including order effects. How can researchers capitalize on testing the “same” people across groups but avoid order effects, history, and maturation? Use an exciting compromise: Match pairs of participants first, making them equal on a variable likely to influence your DV. For example, suppose we wanted to examine the potential effect on heart rate of singing a song in front of an audience versus singing alone. The two IV levels would be singing in front of an audience and singing in a room alone, and heart rate would be measured after the song for both conditions. What additional variable might influence cardiovascular responses to singing alone or in front of people? You probably can think of several variables, such as singing ability. After all, someone who is a great singer should be less nervous than someone who cannot carry a tune.

After you have identified a variable that might impact your study but is not the IV or DV, measure participants on that variable. In the singing example, measure singing ability in your sample. Then match up members of your sample into pairs of participants with similar singing ability. Next, randomly assign one member of the pair to sing in front of an audience and ask the other person to sing alone. Not surprisingly, this method is called **matching** because you are matching people on singing ability and then randomly assigning members of the pair to each IV level. Matching is a behind-the-scenes way for you to create a study with less variability based on individual differences, increasing the chance that you will discover a group differences if it exists.

In most cases, participants in the sample are matched for singing ability before they are randomly assigned to experimental conditions. However, it would be perfectly reasonable to randomly assign all participants to conditions and measure their singing ability at the end of the study. Then you could look over the values for singing ability and match people who are similar to each other, making sure one was in the audience group and the other was in the sing-alone group. You might imagine that matching after participants have already participated in your study will force you to drop some people from your study if they have no match at all (maybe they are professional singers or have absolutely no talent). Or perhaps a participant’s best match on singing ability was in the same experimental condition. As the researcher, you must decide whether to match people on a potentially important variable before the study or after.

Matching.

Matching, or pairing people based on a characteristic you believe will affect the DV, is a behind-the-scenes way to reduce variability, associated with individual differences.



Matching in a Two-Condition Study

We can consider a new example for practice. For this chapter, we will continue our focus on a two-condition study. Recent evidence suggests that taking class notes by hand enhances student test performance over taking notes on a laptop (Mueller & Oppenheimer, 2014). The researchers suggested that students who write out their notes by hand during lecture must process what they are learning to transform the information into fewer words. More effortful thinking about the material results in better retention. We might design a study to test their explanation. If writing with the dominant hand is slower than writing on a computer and thus requires thought and summarizing, writing with the nondominant hand should be even more restrictive. We could ask a new research question: *Do people recall more information when taking notes with their nondominant hand than with their dominant hand?* In the form of a research hypothesis, we would write, *People recall more information when taking notes with their nondominant hand than with their dominant hand.*

The variables must be operationalized, with taking notes using the dominant and nondominant hand offering two levels. Because we can manipulate students' actions by asking them to use either hand, the two conditions represent a true IV, and we can examine cause and effect. The dependent variable is recall of information, which we could measure using a typical test of lecture information. The test will contain 10 multiple-choice items

with applied questions rather than simple factual ones because applied questions better assess deeper processing (learning) of the material.

The Matching Process

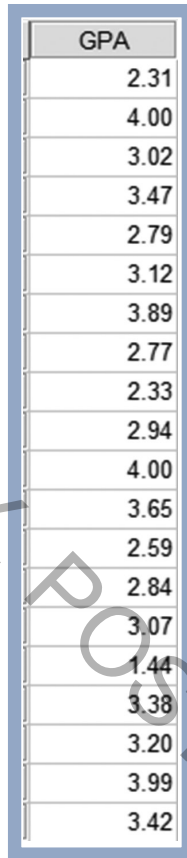
To first match participants in a meaningful way, we have to consider which individual difference would likely influence the DV of test scores. That is, which characteristic of students might introduce nuisance variability to a study of learning? One possibility is IQ, but if we do not have the time or resources to administer a valid and reliable IQ test, we might use self-reported college GPA as an indicator of intelligence. We could send an email to everyone signed up to participate in our study and ask them to report their college GPA. Then we decide which pairs of students best match on GPA. Matching will not be perfect, but we can match any two GPAs that are reasonably similar. Continuing with our behind-the-scenes preparation, we could randomly assign one person in each pair to one of the two conditions and put the remaining member of the pair in the other condition. When participants show up for our study, we would test them in their assigned condition.

Matching in this way requires keeping track of which two people should be paired. Be careful to avoid placing names on data without IRB approval. Consider using a code to keep track of data. In this example, you might ask students to always write the following code when submitting information to you: *Two-digit birthday month + Number of siblings + First four letters of mother's maiden name*. Of course the IRB must approve all parts of your study, including the code you would like to use. In this example of e-mailing for GPA prior to the study, participants can put their GPA on a Word document along with their code and attach it to the e-mail. Then when they arrive for your study, they can provide their code so you can put them in the correct IV level. Again, the IRB will need to approve any process you consider.

After we have GPAs for each participant, we can match participants into pairs. The following data are GPAs to be matched.

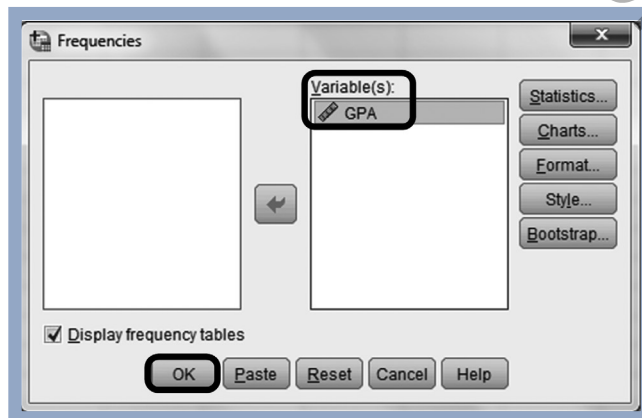
2.31	4.00	3.02	3.47	2.79	3.12	3.89	2.77	2.33	2.94
4.00	3.65	2.59	2.84	3.07	1.44	3.38	3.20	3.99	3.42

Notice that we have 20 participants in the sample. With an even number, we have a good chance of matching everyone into pairs. We need to order GPAs from highest to lowest or lowest to highest. Although we could sort these GPAs by looking at them, you will want to allow a computer program to order the values in a larger data set. To order data, enter GPA into Variable View, and then enter GPAs under Data View.



GPA
2.31
4.00
3.02
3.47
2.79
3.12
3.89
2.77
2.33
2.94
4.00
3.65
2.59
2.84
3.07
1.44
3.38
3.20
3.99
3.42

Click Analyze, Descriptive Statistics, Frequencies. In the box that opens, move GPA to the right using the center arrow, and then click OK.



The output shows a table with GPA frequencies and other values. For our purposes, we need the first column of ordered values circled below.

Frequencies					
GPA					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.44	1	5.0	5.0	5.0
	2.31	1	5.0	5.0	10.0
	2.33	1	5.0	5.0	15.0
	2.59	1	5.0	5.0	20.0
	2.77	1	5.0	5.0	25.0
	2.79	1	5.0	5.0	30.0
	2.84	1	5.0	5.0	35.0
	2.94	1	5.0	5.0	40.0
	3.02	1	5.0	5.0	45.0
	3.07	1	5.0	5.0	50.0
	3.12	1	5.0	5.0	55.0
	3.20	1	5.0	5.0	60.0
	3.38	1	5.0	5.0	65.0
	3.42	1	5.0	5.0	70.0
	3.47	1	5.0	5.0	75.0
	3.65	1	5.0	5.0	80.0
	3.89	1	5.0	5.0	85.0
	3.99	1	5.0	5.0	90.0
	4.00	2	10.0	10.0	100.0
Total		20	100.0	100.0	

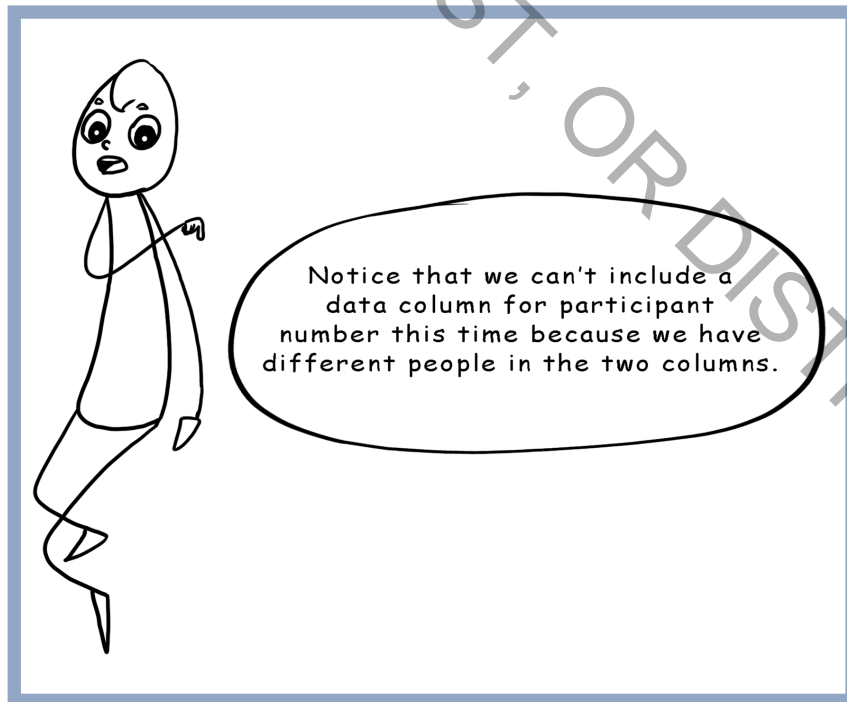
Notice that 4.00 occurs twice, as indicated by the 2 in the Frequency column, which creates a great match between the two students. Those two participants will be in different conditions. Randomly assign the first person to an IV level, and place the second person in the remaining IV level. Next we will pair students with GPAs of 3.99 and 3.89, and so on, randomly assigning the first person of each pair to an IV level and placing the second person in the remaining condition.

	Nondominant Hand	Dominant Hand
GPA Values Paired	4.00	4.00
	3.89	3.99
	3.65	3.47
	3.42	3.38
	3.12	3.20

Nondominant Hand	Dominant Hand
3.02	3.07
2.94	2.84
2.77	2.79
2.59	2.33
1.44	2.31

You might argue that the final pair of students do not have similar GPAs. Use your best judgment. If 1.44 and 2.31 are too different in your mind, you may choose not to use their data in the final analysis.

When a participant arrives for your study, simply look up the code associated with a specific GPA, and test the participant in the assigned IV level. After testing all participants, GPA values in the table are replaced with test scores, the DV of interest. Analyze the data using the paired-samples t -test because the participants have, in fact, been paired.



Nondominant Hand	Dominant Hand
95	98
97	95
92	89
84	80
90	85
82	76
81	80
65	50
71	67
48	43

Again note that our final data set contains the DV of test scores. The GPA values we used to pair students have served their purpose and no longer appear. In a clean SPSS file, enter test scores.

SPSS: Matched-Pairs *t*-test

Under Variable View, label columns with IV levels. Click to Data View to enter data for a paired-samples *t*-test. Click Analyze, Compare Means, Paired-Samples T Test. In the box that opens, move both variables to the right side. If you do not recall the steps, refer to prior examples in this chapter.

SPSS output reveals descriptive statistics in the first table.

T-Test					
Paired Samples Statistics					
		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	NonDominant	80.5000	10	15.29887	4.83793
	Dominant	76.3000	10	18.17232	5.74659

And both the *t*-test outcome and the 95% confidence interval appear in the Paired Samples Test table.

Paired Samples Test					
Paired Differences					
Std. Error Mean	95% Confidence Interval of the Difference		t	df	Sig. (2-tailed)
	Lower	Upper			
1.45144	.91662	7.48338	2.894	9	.018

Based on a Sig value of .018 in the Paired Samples Test table, we know the two groups were meaningfully different. Taking notes with the dominant hand caused significantly different test scores than taking notes with the nondominant hand. We reject the null hypothesis: *People recall the same amount of information when taking notes with their nondominant hand than with their dominant hand.* We revealed support for the research hypothesis: *People recall more information when taking notes with their nondominant hand than with their dominant hand.*

This outcome is good news, but it gets even better. We hypothesized that taking notes with the nondominant hand would require more effortful processing than taking notes with the dominant hand, resulting in higher test scores. As you know, when we clearly expect one group to be higher than the other, we have a directional (one-tailed) test with more power than a nondirectional (two-tailed) test. We divide the nondirectional p -value of .018 by 2 to get a directional p -value of .009.

Now that we know the two groups differed, calculate effect size using Cohen's d . Divide the difference between means (4.20) by the standard deviation (4.59), pulling those two numbers from your SPSS output.

Paired Samples Test					
		Paired Differences			
		Mean	Std. Deviation	Std. Error Mean	95% Co
					Lower
Pair 1	NonDominant - Dominant	4.20000	4.58984	1.45144	

Cohen's d for this example is .92, a large effect. Finally, look at the Paired Samples Statistics table for means to know which group earned higher test scores.

APA Style for the Matched-Pairs t -test

We are ready to create an APA-style results section from the output details. First we provide a method section as an example.

Method

Participants

Twenty college students from a men's college in Denmark participated in this study. Nineteen men reported their ethnicity as White, and 1 reported his ethnicity as Black, with a mean age of 19.78 years ($SD = 1.22$).

Procedure

Prior to the study, the researcher e-mailed participants to request their GPA and asked them to provide a unique code based on their mother's maiden name and number of siblings. The code allowed matching of subsequent data from the same participant without compromising anonymity. Researchers matched pairs of participants based on GPA and randomly assigned each person in a pair to each condition. Students participated individually. When they arrived for the study, participants gave the researcher their unique code and watched a 15-min lecture about pineapples while taking notes with either their dominant or nondominant hand. Performance on 10 multiple-choice, applied items constituted the dependent variable.

Results

We analyzed these data using a one-tailed, matched-pairs t -test. Taking notes with the dominant versus nondominant hand affected test scores, $t(9) = 2.89, p = .009, 95\% \text{ CI } [0.92, 7.48], d = .92$. Students who wrote lecture notes with their nondominant hand earned higher test scores ($M = 80.50, SD = 15.30, n = 10$) than those who took notes with their dominant hand ($M = 76.30, SD = 18.17, n = 10$).

Note that we first measured participants on college GPA to pair people by similar GPAs as an indication of similar intelligence. We matched for GPA because the DV was test grades, which reasonably could be influenced by intelligence. Matching first for GPA gave us a better chance of finding an effect of note-taking approaches on test grades. Pairing participants based on a relevant variable offers more statistical power, and you will be more likely to find group differences on your DV if a difference truly exists.

SUMMARY

In this chapter, we explained two-group designs in which either the same people were tested twice or participants were first matched in some way. These designs reduce individual differences and increase the chance of finding a significant study outcome. When the same people are tested twice, several order effects can occur, requiring a solid research design using counterbalancing of conditions. If participants are manipulated with a true IV,

researchers can establish cause and effect. If, on the other hand, a quasi-IV is used, we can examine a potential relationship between the quasi-IV and DV. Regardless of whether an IV or quasi-IV is chosen, related-samples designs with two levels are analyzed using a *t*-test.

REVIEW OF TERMS

Counterbalancing	Order Effects (Carryover Effects)
Fatigue Effect	Practice Effect
History Effect	Related-Samples (Dependent-Samples, Paired-Samples, Matched-Pairs) Design
Matching	Related-Samples <i>t</i> -test
Maturation	

PRACTICE ITEMS

1. What is the difference between a related-samples design and an independent-samples design?
2. Why are related-samples designs considered to be more “powerful” than independent-samples designs?
3. Discuss several problems with testing the same people twice and how you might “solve” those problems.
4. When conducting a study, why might we match participants on a variable rather than have different people in levels of our IV?
5. What effect-size term is associated with the paired-samples *t*-test, and what is considered a weak, moderate, and strong effect?

* * *

For each of the following studies, (a) restate the research question as a research hypothesis and state the null hypothesis, (b) determine how many participants are needed for adequate power, and (c) enter and analyze the data as well as write an APA-style results section. We have written method sections for you as examples.

6. One night when you are in a study group preparing for a final exam, you notice that several of your classmates spend time checking social media pages instead of studying. You read in Panek (2014) that use of social media is related to productivity, with more social media use related to lower productivity. You wonder if access to technology, in general, reduces how prepared students feel after an exam study

session. Based on the Panek study, you think students might feel less prepared. But based on the possibility that cell phones can be used to look up answers to study questions, students might feel more prepared. You ask the research question: *If students have access to their cell phones while studying, will they feel more or less prepared for the exam?* Below, data are presented as pairs of participants with similar GPAs. Higher values for preparation indicate students feel more prepared.

Access to Cell Phones	No Access to Cell Phones
1	0
1	3
2	0
2	4
2	2
3	1
3	1
4	3
4	4
5	0
6	8
7	6
7	5
8	4
8	4
9	5
9	7
10	9
10	10
10	8

Method

Participants

We recruited students ($N = 40$; 50% women, 50% men) in an Introduction to Anthropology class using a flyer posted outside the classroom. Ethnicities included 28 Latino, 5 Black, 5 White, and 2 Chinese individuals, with a mean age of 19.20 ($SD = 2.58$). Researchers entered all participants into a drawing for \$25.

Procedure

Students arrived at the study session, which was held in the classroom next door to their regular classroom, 1 hr before a scheduled class exam. At the beginning of the study session, students reported their GPA, and researchers matched them based on this variable. Within each pair, the researchers randomly assigned students to treatment conditions and distributed one of two sheets of paper to each participant. One paper instructed students in the first condition to leave their phones on the table during the study session, and the other paper instructed students in the second condition to turn their phones off and place them under the table. At the end of the study session, students rated how prepared they felt for the exam using a scale of 0 (*not prepared at all*) to 10 (*extremely prepared*).

7. You have a daughter with autism spectrum disorder (ASD), which has sparked your interest in treatments for this disorder. Your specific interest is in animal-assisted therapies for ASD. You read a study by Ward, Whalon, Rusnak, Wendell, and Paschall (2014) finding that elementary-age children with ASD who engaged in therapeutic horseback riding scored lower on teacher ratings of ASD impairment. (Higher ratings indicate more impairment.) You wonder if, among older teens with ASD, the same intervention might help reduce self-reported impairment associated with ASD. You ask the research question: *Does horseback riding reduce impairment among high-school-age teens with ASD?*

(data continued)

Participant Number	Pretest	Posttest
1	7	4
2	3	1
3	1	2
4	3	1
5	5	5
6	5	4
7	3	2
8	7	3

Participant Number	Pretest	Posttest
9	2	3
10	3	1
11	4	2
12	6	1
13	4	4
14	6	5
15	4	4
16	2	5

Method

Participants

Participants included 16 teenagers (10 boys, 5 girls, and 1 nonbinary individual) with parent-reported previous diagnoses of autism spectrum disorder (ASD). Ages ranged from 15 to 17 ($M = 16.51$, $SD = 0.97$), and ethnicities included 14 White, 1 Black, and 1 mixed-race individual. In addition to parent consent, researchers obtained child assent prior to both pre- and posttest data collection.

Procedure

Participants reported their impairment on a scale of 1 (*very little impairment*) to 7 (*a great deal of impairment*) before and after a 6-week horseback riding intervention. Although all 16 participants participated in the intervention as a group, each person rode his or her own horse with an instructor present. Participants rode the same horse and worked with the same instructor at each weekly 1-hr session.

8. Hancock, Jorgensen, and Swanson (2013) found several factors related to credit-card use and debt among college students, concluding that early intervention may be the best way to reduce credit-card debt. You ask, *Will first-year college students exposed to a lecture about finances have less credit-card debt one year later than students not exposed to this lecture?* You recruit pairs of twins, assuming that twins will enter college with similar amounts of debt, so you can use a related-samples design. The data in the table below are dollar amounts of debt presented for pairs of twins.

(data continued)

Lecture	No Lecture
2000	2500
500	600
450	100
150	300
4500	4000
4260	4710
1000	2560

(data continued)

Lecture	No Lecture
2600	1900
10100	12010
460	580
5000	6050
8000	6000
4710	4700
3000	3500

(data continued)

Lecture	No Lecture
4000	1000
2000	2500
750	1500
6580	7890
2580	2620
2500	8000
6050	5020

Method

Participants

We recruited 21 sets of twins (12 fraternal, 9 identical) in their first semester of college. Of these, 23 individuals identified as female gender, 18 identified as male gender, and 1 chose not to provide information about gender. Ethnicities included 22 Black, 16 White, and 4 Latino individuals, with a mean age of 18.76 ($SD = 1.01$). Sexual orientations included 32 straight, 8 gay, and 2 pansexual individuals.

Procedure

We recruited participants from a large university in Colorado via posts on Facebook and Twitter. As twins arrived for the study, we randomly assigned them to either the treatment or control group. Those in the treatment group watched a 10-min video lecture about credit-card debt in one room, while those in the control group watched a 10-min video about staying safe on campus in another room. At the end of the spring semester, we asked twins to e-mail their total amount of credit-card debt.

REFERENCES

- Bascandziev, I., & Harris, P. L. (2014). In beauty we trust: Children prefer information from more attractive informants. *British Journal of Developmental Psychology, 32*, 94–99. doi:10.1111/bjdp.12022
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.
- Hancock, A. M., Jorgensen, B. L., & Swanson, M. S. (2013). College students and credit card use: The role of parents, work experience, financial knowledge, and credit card attitudes. *Journal of Family and Economic Issues, 34*, 369–381. doi:10.1007/s10834-012-9338-8
- Hively, K., & El-Alavli, A. (2014). “You throw like a girl”: The effect of stereotype threat on women’s athletic performance and gender stereotypes. *Psychology of Sport and Exercise, 15*(1), 48–55. doi:10.1016/j.psychsport.2013.09.001
- Mueller, P. A., & Oppenheimer, D. M. (2014). The pen is mightier than the keyboard: Advantages of longhand over laptop note taking. *Psychological Science, 25*(6), 1159–1168. doi:10.1177/0956797614524581
- Panek, E. (2014). Left to their own devices: College students’ “guilty pleasure” media use and time management. *Communication Research, 41*(4), 561–577. doi:10.1177/0093650213499657
- Ward, S. C., Whalon, K., Rusnak, K., Wendell, K., & Paschall, N. (2014). The association between therapeutic horseback riding and the social communication and sensory reactions of children with autism. *Journal of Autism and Developmental Disorders, 43*, 2190–2198. doi:10.1007/s10803-013-1773-3