CHAPTER THIRTEEN

Data Analysis and Interpretation: Part II. Tests of Statistical Significance and the Analysis Story

CHAPTER OUTLINE

OVERVIEW

In Chapter 12 we introduced the three major stages of data analysis: *getting to know the data, summarizing the data, and confirming what the data tell us*. In the final stage of data analysis we evaluate whether we have sufficient evidence to make a claim about behavior. What, given these data, can we say about behavior? This stage is sometimes called *confirmatory data analysis* (e.g., Tukey, 1977). At this point we seek confirmation for what the data are telling us. In Chapter 12 we emphasized the use of confidence intervals to confirm what the data tell us. In this chapter we continue our discussion of confirmatory data analysis by focusing on tests of statistical significance, or what is more formally known as *null hypothesis significance testing* (NHST).

NHST is the most common approach to performing confirmatory data analysis. Nevertheless, tests of statistical significance have received persistent criticism (e.g., Cohen, 1995; Hunter, 1997; Loftus, 1991, 1996; Meehl, 1967; Schmidt, 1996), and for good reason. Researchers have been misusing (and misinterpreting) them for decades, all the time ignoring warnings that they were doing so (e.g., Finch, Thomason, & Cumming, 2002). There are critics who suggest we discard NHST altogether (e.g., Hunter, 1997; Schmidt, 1996). For example, an alternative approach focuses not on significance testing but on the probability of replicating an effect. This statistic, noted as p_{rep}, can be computed whenever an effect size can be calculated (see Killeen, 2005). However, the majority of experts suggest that we continue to use NHST but be cautious about its use (e.g., Abelson, 1995, 1997; Chow, 1988; Estes, 1997; Greenwald, Gonzalez, Harris, & Guthrie, 1996; Hagen, 1997; Krueger, 2001; Mulaik, Raju, & Harshman, 1997). Whatever the outcome of this debate within the psychology community, there is nearly universal agreement on the need (a) to understand exactly what it is that NHST can and cannot do, and (b) to increase our use of alternative methods of data analysis, especially the use of confidence intervals and the reporting of effect sizes. Sometimes these alternative techniques will supplant NHST, at other times they will complement NHST.

In what immediately follows we first provide an overview of NHST. Next we discuss the important concepts of experimental sensitivity and statistical power. Then we illustrate the NHST approach to data analysis using the same data we used in Chapter 12 to construct confidence intervals for the difference between two means. By using the same data, we can contrast the information obtained from NHST with that provided by confidence intervals. We point out what we can and cannot say based on NHST and suggest that information obtained from NHST can complement information obtained with confidence intervals. Finally, we provide some recommendations for you to follow when evaluating evidence for a claim about behavior involving two means and illustrate how to create an analysis story for your study.

The most common technique of confirmatory data analysis associated with studies involving more than two groups is a form of NHST called *analysis of variance (ANOVA)*. The rationale for using an ANOVA, the computational procedures associated with ANOVA, and the interpretation of ANOVA results are discussed in the second half of this chapter.

Null Hypothesis Significance Testing (NHST)

- Null hypothesis testing is used to determine whether mean differences among groups in an experiment are greater than the differences that are expected simply because of error variation.
- The first step in null hypothesis testing is to assume that the groups do not differ—that is, that the independent variable did not have an effect (the null hypothesis).
- Probability theory is used to estimate the likelihood of the experiment's observed outcome, assuming the null hypothesis is true.
- A statistically significant outcome is one that has a small likelihood of occurring if the null hypothesis were true.
- Because decisions about the outcome of an experiment are based on probabilities, Type I (rejecting a true null hypothesis) or Type II (failing to reject a false null hypothesis) errors may occur.

Statistical inference is both inductive and indirect. It is inductive because we draw general conclusions about populations on the basis of the specific samples we test in our experiments, as we do when constructing confidence intervals. However, unlike the approach using confidence intervals, this form of statistical inference is also indirect because it begins by assuming the null hypothesis. The **null hypothesis (H_0)** is the assumption that the independent variable has had no effect. Once we make this assumption, we can use probability theory to determine the likelihood of obtaining this difference (or a larger difference) observed in our experiment *IF* the null hypothesis were true. If this likelihood is small, we reject the null hypothesis and conclude that the independent variable did have an effect on the dependent variable. Outcomes that lead us to reject the null hypothesis are said to be *statistically significant*. A statistically significant outcome means only that the difference we obtained in our experiment is larger than would be expected if error variation alone (i.e., chance) were responsible for the outcome (see Box 13.1).

Key Concept

A statistically significant outcome is one that has only a small likelihood of occurring if the null hypothesis were true. But just how small is small enough? Although there is no definitive answer to this important question, the consensus among members of the scientific community is that outcomes associated with probabilities of less than 5 times out of 100 (or .05) if the null hypothesis were true are judged to be statistically significant. The probability we elect to use to indicate an outcome is statistically significant is called the **level of significance.** The level of significance is indicated by the Greek letter alpha (α). Thus, we speak of the .05 level of significance, the .10 level, or the .01 level, which we report as $\alpha = .05$, $\alpha = .10$, $\alpha = .01$, respectively.

Key Concept

Just what do our results tell us when they are statistically significant? The most useful information we gain is that we know that something interesting has happened. More specifically, we know that the smaller the exact probability of the observed outcome, the greater is the probability that an exact replication will produce a statistically significant finding. But we must be careful what we mean by this statement. Researchers sometimes mistakenly say that when a result occurs with $p < .05$, "This outcome will be obtained 95/100 times if the

BOX 13.1

HEADS OR TAILS? TOSSING COINS AND NULL HYPOTHESES

Perhaps you can appreciate the process of statistical inference by considering the following dilemma. A friend, with a sly smile, offers to toss a coin with you to see who pays for the meal you just enjoyed at a restaurant. Your friend just happens to have a coin ready to toss. Now it would be convenient if you could directly test whether your friend's coin is biased (by asking to look at it). Not willing to appear untrusting, however, the best you can do is test your friend's coin indirectly by assuming it is not biased and seeing if you consistently get outcomes that differ from the expected 50:50 split of heads and tails. If the coin does not exhibit the ordinary 50:50 split (after many trials of flipping the coin), you might surmise that your friend is trying, by slightly underhanded means, to get you to pay for the meal. Similarly, we would like to make a direct test of statistical significance for an obtained outcome in our experiments. The best we can do, however, is to compare our obtained outcome with the expected outcome of no difference between frequencies of heads and tails. *The key to understanding null hypothesis testing is to recognize that we can use the laws of probability to estimate the likelihood of an outcome only when we assume that chance factors are the sole cause of that outcome.* This is not different from flipping your friend's coin a number of times to make your conclusion. You know that, based on chance alone, 50% of the time the coin should come up heads, and 50% of the time it should be tails. After many coin tosses, anything different from this probable outcome would lead you to conclude that something other than chance is working—that is, your friend's coin is biased.

study is repeated." This is simply not true. Simply achieving statistical significance (i.e., $p < .05$) does not tell us about the probability of replicating the results. For example, a result just below .05 probability (and thus statistically significant) has only about a 50:50 chance of being statistically significant (i.e., $p < .05$) if replicated exactly (Greenwald et al., 1996). On the other hand, knowing the exact probability of the results does convey information about what will happen if a replication were done. The smaller the exact probability of an initial finding, the greater the probability that an exact replication will produce a statistically significant ($p < .05$) finding (e.g., Posavac, 2002). Consequently, and as recommended by the American Psychological Association (APA), *always report the exact probability of results when carrying out NHST.*

You must choose the level of significance before you begin your experiment, not after you have done the statistical analysis. Choosing the level of significance before doing the analysis allows you to avoid the temptation of using the probability of your obtained outcome as the level of significance you would have chosen. Strictly speaking, there are only two conclusions possible when you do an inferential statistics test: Either you *reject* the null hypothesis or you *fail to reject* the null hypothesis. Note that we did *not* say that one alternative is to accept the null hypothesis. Let us explain.

When we conduct an experiment and observe the effect of the independent variable is not statistically significant, we do not reject the null hypothesis. However, neither do we necessarily accept the null hypothesis of no difference. There may have been some factor in our experiment that prevented us from

observing an effect of the independent variable (e.g., ambiguous instructions to subjects, poor operationalization of the independent variable). As we will show later, too small a sample often is a major reason why a null hypothesis is not rejected. Although we recognize the logical impossibility of proving that a null hypothesis is true, we also must have some method of deciding which independent variables are not worth pursuing. NHST can help with that decision. A result that is not statistically significant suggests we should be cautious about concluding that the independent variable influenced behavior in more than a trivial way. At this point you will want to seek more information, for example, by noting the size of the sample and the effect size (see the next section, "Experimental Sensitivity and Statistical Power").

There is a troublesome aspect to the process of statistical inference and our reliance on probabilities for making decisions. No matter what decision you reach, and no matter how carefully you reach it, there is always some chance you are making an error. The two possible "states of the world" and the two possible decisions an experimenter can reach are listed in Table 13.1. The two "states of the world" are that the independent variable either does or does not have an effect on behavior. The two possible correct decisions the researcher can make are represented by the upper-left and lower-right cells of the table. If the independent variable does have an effect, the researcher should reject the null hypothesis; if it does not, the researcher should fail to reject the null hypothesis.

The two potential errors (Type I error and Type II error) are represented by the other two cells of Table 13.1. These errors arise because of the probabilistic nature of statistical inference. When we decide an outcome is statistically significant because the outcome's probability of occurring under the null hypothesis is less than .05, we acknowledge that in 5 out of every 100 tests, the outcome could occur even if the null hypothesis were true. The level of significance, therefore, represents the probability of making a **Type I error:** rejecting the null hypothesis when it is true. The probability of making a Type I error can be reduced simply by making the level of significance more stringent, perhaps .01. The problem with this approach is that it increases the likelihood of making a **Type II error:** failing to reject the null hypothesis when it is false.

Key Concepts

The problem of Type I errors and Type II errors should not immobilize us, but it should help us understand why researchers rarely use the word "prove" when they describe the results of an experiment that involved tests of statistical significance. Instead, they describe the results as "consistent with the hypothesis," or "confirming the hypothesis," or "supporting the hypothesis." These tentative

TABLE 13.1 POSSIBLE OUTCOMES OF DECISION MAKING WITH INFERENTIAL STATISTICS

	States of the world	
	Null hypothesis is false.	**Null hypothesis is true.**
Reject null hypothesis	Correct decision	Type I error
Fail to reject null hypothesis	Type II error	Correct decision

statements are a way of indirectly acknowledging that the possibility of making a Type I error or a Type II error always exists. The .05 level of significance represents a compromise position that allows us to strike a balance and avoid making too many of either type of error. The problem of Type I errors and Type II errors also reminds us that *statistical inference can never replace replication as the best test of the reliability of an experimental outcome.*

Experimental Sensitivity and Statistical Power

- Sensitivity refers to the likelihood that an experiment will detect the effect of an independent variable when, in fact, the independent variable truly has an effect.
- Power refers to the likelihood that a statistical test will allow researchers to reject correctly the null hypothesis of no group differences.
- The power of statistical tests is influenced by the level of statistical significance, the size of the treatment effect, and the sample size.
- The primary way for researchers to increase statistical power is to increase sample size.
- Repeated measures designs are likely to be more sensitive and to have more statistical power than independent groups designs because estimates of error variation are likely to be smaller in repeated measures designs.
- Type II errors are more common in psychological research using NHST than are Type I errors.
- When results are not statistically significant (i.e., $p > .05$), it is incorrect to conclude that the null hypothesis is true.

Key Concept

The *sensitivity of an experiment* is the likelihood that it will detect an effect of the independent variable if the independent variable does, indeed, have an effect (see Chapter 8). An experiment is said to have sensitivity; a statistical test is said to have **power.** The power of a statistical test is the probability that the null hypothesis will be rejected when it is false. The null hypothesis is the hypothesis of "no difference" and, thus, is false and should be rejected when the independent variable has made a difference. Recall that we defined a Type II error as the probability of failing to reject the null hypothesis when it is false. Power can also be defined as 1 minus the probability of a Type II error.

Power tells us how likely we are to "see" an effect that is there and is an estimate of the study's replicability. Because power tells us the probability of rejecting a false null hypothesis, we know how likely we are to miss a real effect. For instance, if a result is not significant and power is only .30, we know that a study with these characteristics detects an effect equal to the size we observed only 3 out of 10 times. Therefore, 7 of 10 times we do this study we will miss seeing the effect. In this case we may want to suspend judgment until the study can be redone with greater power.

The power of a statistical test is determined by the interplay of three factors: the level of statistical significance, the size of the treatment effect, and the sample size (Keppel, 1991). For all practical purposes, however, *sample size is the primary factor that researchers use to control power*. The differences in sample size

that are needed to detect effects of different sizes can be dramatic. For example, Cohen (1988) reports the sample sizes needed for an independent groups design experiment with one independent variable manipulated at three levels. It takes a sample size of 30 to detect a large treatment effect; it takes a sample size of 76 to detect a medium treatment effect; and it takes a sample size of 464 to detect a small treatment effect. It thus takes over 15 times more participants to detect a small effect than it does to detect a large effect!

Using repeated measures experiments can also affect the power of the statistical analyses researchers use. As described in Chapter 8, repeated measures experiments are generally more sensitive than are independent groups experiments. This is because the estimates of error variation are generally smaller in repeated measures experiments. The smaller error variation leads to an increased ability to detect small treatment effects in an experiment. And that is just what the power of a statistical analysis is—the ability to detect small treatment effects when they are present.

When introducing NHST we suggested that making a so-called Type I error is equivalent to alpha (.05 in this case). Logically, to make this kind of error, the null hypothesis must be capable of being false. Yet, critics argue that the null hypothesis defined as zero difference is "always false" (e.g., Cohen, 1995, p. 1000) or, somewhat more conservatively, is "rarely true" (Hunter, 1997, p. 5). If an effect is always, or nearly always, present (i.e., there is more than a zero difference between means), then we can't possibly (or at least hardly ever) make a mistake by claiming that an effect is there when it is not. Following this line of reasoning, the only error we are capable of making is a Type II error (see Hunter, 1997; Schmidt & Hunter, 1997), that is, saying a real effect is not there. This type of error, largely due to low statistical power in many psychological studies, typically is much greater than .05 (e.g., Cohen, 1990; Hunter, 1997; Schmidt & Hunter, 1997). Let us suggest that Type I errors do occur if the null hypothesis is taken literally, that is, if there really is a literally zero difference between the population means or if we believe that in some situations it is worth testing an effect against a hypothesis of no difference (see Abelson, 1997; Mulaik et al., 1997). As researchers we must be alert to the fact that in some situations it may be important not to conclude an effect is present when it is not, at least not to more than a trivial degree (see Box 13.2).

Type II errors are likely when power is low, and low power has characterized many studies in the literature: *The most common error in psychological research using NHST is a Type II error*. Just because we did not obtain statistical significance does not mean that an effect is not present (e.g., Schmidt, 1996). In fact, one important reason for obtaining a measure of effect size is that we can compare the obtained effect with that found in other studies, whether or not the effect was statistically significant. This is the goal of meta-analysis (see Chapter 7). Although a nonsignificant finding does not tell us that an effect is absent, assuming that our study was conducted with sufficient power, a nonsignificant finding may indicate that an effect is so small that it isn't worth worrying about.

To determine the power of your study *before* it is conducted, you must first estimate the effect size anticipated in your experiment. An examination of the

BOX 13.2

DO WE EVER ACCEPT THE NULL HYPOTHESIS?

Despite what we have said thus far, there may be some instances in which researchers will choose to accept the null hypothesis (rather than simply fail to reject it). Yeaton and Sechrest (1986, pp. 836–837) argue persuasively that findings of no difference are especially critical in applied research. Consider some questions they cite to illustrate their point: Are children who are placed in daycare centers as intellectually, socially, and emotionally advanced as children who remain in the home? Is a new, cheaper drug with fewer side effects as effective as the existing standard in preventing heart attacks?

These important questions clearly illustrate situations in which accepting the null hypothesis (no effect) involves more than a theoretical issue—life and death consequences rest on making the correct decision. Frick (1995) argues that never accepting the null hypothesis is neither desirable nor practical for psychology. There may be occasions when we want to be able to state with confidence that there is no (meaningful) difference (see also Shadish, Cook, & Campbell, 2002).

effect sizes obtained in previous studies for the independent variable of interest should guide your estimate. Once an effect size is estimated, you must then turn to "power tables" to obtain information about the sample size you should use in order to "see" the effect. These steps for conducting a power analysis are described more fully in the Appendix, where a power table for comparing two means is found. *When you have a good estimate of the effect size you are testing, it is strongly recommended that you perform a power analysis before doing a research study*.

Power tables are also used after the fact. When a study is completed and the finding is not statistically significant, the APA *Publication Manual* (2001) recommends that the power of your study be reported. In this way you communicate to other researchers the likelihood of detecting an effect that was there. If that likelihood was low, then the research community may wish to suspend judgment regarding the meaning of your findings until a more powerful replication of your study is carried out. On the other hand, a statistically nonsignificant result from a study with sufficient power may suggest to the research community that this is an effect not worth pursuing. Instructions for doing this type of power analysis are also found in the Appendix.

NHST: Comparing Two Means

- The appropriate inferential test when comparing two means obtained from different groups of subjects is a *t*-test for independent groups.
- A measure of effect size should always be reported when NHST is used.
- The appropriate inferential test when comparing two means obtained from the same subjects (or matched groups) is a repeated measures (within-subjects) *t*-test.

We now illustrate the use of NHST when comparing the difference between two means. First, we consider a research study involving two independent means. The data for this study are from our example vocabulary study, which we described in Chapter 12. Next we consider a situation where there are two dependent means, that is, when a repeated measures design was used.

Independent Groups

Key Concept

Recall that a study was conducted in which the vocabulary size of college students and older adults was assessed. The appropriate inferential test for this situation is a ***t*-test for independent groups.** We may use this test to evaluate the difference between the mean percent multiple-choice performance of the college and older adult samples. We can define t for independent groups as the difference between sample means $(\overline{X}_1 - \overline{X}_2)$ divided by the standard error of the mean difference $(s_{\overline{X}_1 - \overline{X}_2})$. That is,

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{\overline{X}_1 - \overline{X}_2}} \quad \text{where } s_{\overline{X}_1 - \overline{X}_2} = \sqrt{\left[\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}\right]\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}$$

Using this formula, the obtained t was 5.84. An alpha of .05 had been chosen prior to doing the analysis. Appendix Table A.2 shows critical values of t associated with various degrees of freedom (df). The obtained t of 5.84 is larger than the critical t value with 50 df (the $df = N - 2$, where N equals the sum of the two sample sizes, $n_1 + n_2$); thus, the obtained t can be said to be statistically significant.[1]

Statistical software programs typically provide the actual probability of an obtained t as part of the output, thus circumventing the need to consult a table of t values. In fact, the APA *Publication Manual* (2001) advises that the exact probability be reported. When the exact probability is less than .001 (e.g., $p = .0004$), statistical software programs frequently report the exact probability as .000. (This was the case for the analysis reported above.) Of course, the exact probability is not .000 but something less than .001. Because in this situation a researcher does not know exactly what that probability is, one solution is to report the obtained probability as $p \leq .0005$ since a value larger than .0005 would have been rounded to .001.

Therefore, for the vocabulary study we have been discussing, the result of the inferential statistics test can be summarized as

$$t(50) = 5.84, p \leq .0005$$

[1]The critical t value for 50 df is not found in Appendix Table A.2. When this occurs it is appropriate to compare the obtained t with the critical t associated with the next lowest number of degrees of freedom. In Table A.2, the value associated with 40 df is 2.02. The obtained value of 5.84 is clearly larger than this value. As pointed out in the text, when statistical analysis is done using computer software packages, the exact probability is given automatically. This exact probability should always be reported.

In Chapter 12 we showed how an effect size, *d*, can be calculated for a comparison between two means. *A measure of effect size should always be reported when NHST is used*. You may recall that in Chapter 12 we calculated *d* for the vocabulary study as 1.65. Cohen's *d* also can be calculated from the outcome of the independent groups *t*-test according to the following formula:

$$d = \frac{2t}{\sqrt{df}} \quad \text{(see Rosenthal \& Rosnow, 1991)}$$

That is,

$$d = \frac{2(5.84)}{\sqrt{50}} = \frac{11.68}{7.07} = 1.65$$

Repeated Measures Designs

Thus far we have considered experiments involving two independent groups of subjects. As you are aware, experiments can also be carried out by having each subject participate in each condition of the experiment or by "matching" subjects on some measure related to the dependent variable (e.g., IQ scores, weight). Such experiments are called matched groups (see Chapter 7), within-subjects designs, or repeated measures designs (see Chapter 8). The logic of NHST is the same in a repeated measures design as it is in an independent groups design. However, the *t*-test comparing two means takes on a different form in a repeated measures design. The *t*-test in this situation is typically called a direct-difference *t* or **repeated measures (within-subjects) *t*-test.** When each subject is in both conditions of the experiment, *t* is defined as

Key Concept

$$t = \frac{\overline{D}}{s_{\overline{D}}}$$

where

$\overline{D}$ = mean of difference scores or $(\Sigma D)/N$

$s_{\overline{D}}$ = standard error of difference scores, that is,

$$s_{\overline{D}} = s_D/\sqrt{N}$$

where

s_D = standard deviation of difference scores

N = number of difference scores (i.e., number of pairs of scores)

The numerator of the repeated measures *t* is the mean of the difference scores ($\overline{D}$) and is algebraically equivalent to the difference between the sample means (i.e., $\overline{X}_1 - \overline{X}_2$). Statistical significance is determined by consulting a table of critical values of *t* with *df* equal to $N - 1$ (see Appendix Table A.2). In this case, *N* refers to the number of participants or pairs of scores in the experiment. You interpret the obtained *t* as you would the *t* obtained in an independent groups design.

As noted in Chapter 12, assessing effect size in a matched groups or repeated measures design is somewhat more complex than for an independent groups design (see Cohen, 1988, and Rosenthal & Rosnow, 1991, for information pertaining to the calculation of d in these cases).

Statistical Significance and Scientific or Practical Significance

- We must recognize the fact that statistical significance is not the same as scientific significance.
- We also must acknowledge that statistical significance is not the same as practical or clinical significance.

Tests of statistical significance are an important tool in the analysis of research findings. We must be careful, however, to interpret statistically significant findings correctly (see Box 13.3). We must also be careful not to confuse a statistically significant finding with a scientifically significant finding. Whether the results of a study are important to the scientific community will depend on the nature of the variable under study (the effects of some variables are simply more important than those of others), how sound the study is (statistically significant findings can be produced with poorly done studies), and other criteria such as effect size (see, for example, Abelson, 1995).

Similarly, the practical or clinical significance of a treatment effect depends on factors other than statistical significance. These include the external validity associated with the study, the size of the effect, and various practical considerations (including financial ones) associated with a treatment's implementation. Even a statistically significant outcome showing a large effect size is not a guarantee of its practical or clinical significance. A very large effect size might be obtained as a part of a study that does not generalize well from the laboratory to the real world (i.e., has low external validity); thus, the results may be of little value to the applied psychologist. Moreover, a relatively large treatment effect that does generalize well to real-world settings may never be applied because it is too costly, too difficult to implement, too controversial, or too similar in its effects to existing treatments.

It is also possible that, given enough power, a small effect size will be statistically significant. Small effect sizes may not be practically important outside the laboratory. As we described in Chapter 7, external validity is an empirical question. It is important to conduct a study under conditions similar to those in which the treatment will be used in order to see whether a finding is practically significant. We are not likely to carry out such an empirical test, however, if the effect size is small (although see Rosenthal, 1990, for important exceptions).

Recommendations for Comparing Two Means

We offer the following recommendations when evaluating the data from a study looking at the difference between two means. First, keep in mind the final goal of data analysis: to make a case based on our observations for a claim about behavior. In order to make the best case possible, you will want to explore various alternatives for data analysis. Don't fall into the trap of thinking that there

BOX 13.3

WHAT WE SHOULD NOT SAY WHEN A RESULT IS STATISTICALLY SIGNIFICANT ($p < .05$)

- We cannot specify the exact probability for the real difference between the means. For example, it is wrong to say that the probability is .95 that the observed difference between the means reflects a real (true) mean difference in the populations.

The outcome of NHST reveals the probability of a difference this great by chance (given these data) assuming the null hypothesis is true. It does not tell us about probabilities in the real world (e.g., Mulaik et al., 1997). If results occur with a probability less than our chosen alpha level (e.g., .05), then all we can conclude is that the outcome is not likely to be a chance event in this situation.

- Statistically significant results do not demonstrate that the research hypothesis is correct. (For example, the data from the vocabulary study do not prove that older adults have greater vocabulary knowledge than do younger adults.)

NHST (as well as confidence intervals) cannot prove that a research hypothesis is correct. A statistically significant result is (reasonably) sometimes said to "provide support for" or to "give evidence for" a hypothesis, but it alone cannot prove that the research hypothesis is correct. There are a couple of important reasons why. First, NHST is a game of probabilities; it provides answers in the form of likelihoods that are never 1.00 (e.g., p greater or less than .05). There is always the possibility of error. If there is "proof," it is only "circumstantial" proof. As we have seen, the research hypothesis can only be tested indirectly by referring to the probability of these data assuming the null hypothesis is true. If the probability that our results occurred by chance is very low (assuming a true null hypothesis), we may reason that the null hypothesis is really not true; this does not, however, mean our research hypothesis is true. As Schmidt and Hunter (1997, p. 59) remind us, researchers doing NHST "are focusing not on the actual scientific hypothesis of interest." Second, evidence for the effect of an independent variable is only as good as the methodology that produced the effect. The data used in NHST may or may not be from a study that is free of confounds or experimenter errors. It is possible that another factor was responsible for the observed effect. (For example, suppose that the older adults in the vocabulary study, but not the college students, had been recruited from a group of expert crossword puzzle players.) As we have mentioned, a large effect size can easily be produced by a bad experiment. Evidence for a research hypothesis must be sought by examining the methodology of a study as well as considering the effect produced on the dependent variable. *Neither NHST, confidence intervals, nor effect sizes tell us about the soundness of a study's methodology.*

is one and only one way to provide evidence for a claim about behavior. When there is a choice (and there almost always is), as recommended by the APA's Task Force on Statistical Inference (Wilkinson et al., 1999), use the simplest possible analysis. Second, when using NHST be sure to understand its limitations and what the outcome of NHST allows you to say. Always consider reporting a measure of effect magnitude when using NHST, and also a measure of power, especially when a nonsignificant result is found. Although there will be some situations when effect size information is not warranted—for example, when testing a theoretical prediction of direction only (e.g., Chow, 1988), these situations are relatively rare. In many research situations, and in nearly all applied situations, effect size information is an important, even necessary, complement to

NHST. Finally, researchers must "break the habit" of relying solely on NHST and consider reporting confidence intervals for effect sizes in addition to, or even rather than, *p* values associated with results of inferential tests. The APA *Publication Manual* (2001, p. 22) strongly recommends the use of confidence intervals.

Reporting Results when Comparing Two Means

We are now in a position to model a statement of results that takes into account the information gained from all three stages of data analysis, the complementary evidence obtained by using confidence intervals (Chapter 12) and NHST, and the recommendations of the APA *Publication Manual* (2001) regarding reporting results (see especially pp. 20–26 of the *Manual*).

Reporting Results of the Vocabulary Study We may report the results as follows:

> The mean performance on the multiple-choice vocabulary test for college students was 45.58 ($SD = 10.46$); the mean of the older group was 64.04 ($SD = 12.27$). With alpha set at .05, this difference was statistically significant, $t(50) = 5.84$, $p \leq .0005$. Older participants in this study had a greater vocabulary size than did the younger participants. The effect size based on Cohen's *d* was 1.65. There is a .95 probability that the obtained confidence interval, 12.11 to 24.81, contains the true population mean difference.

Commentary Descriptive statistics in the forms of means and standard deviations summarize "what happened" in the experiment as a function of the independent variable (age). As recommended by the APA *Publication Manual,* the alpha level for NHST is stated prior to reporting the obtained *p* value. Because the exact probability was less than .001, results are reported at $p \leq .0005$, but note that exact probabilities are to be reported when .001 or greater. The exact probability conveys information about the probability of an exact replication (Posavac, 2002). That is, we know that the results are "more reliable" than if a larger exact *p* value was obtained. This information is not found when only confidence intervals are reported. The sentence beginning "Older participants in this study . . ." summarizes in words what the statistical analysis revealed. It is always important to tell your reader directly what the analysis shows. This becomes increasingly important as the number and complexity of analyses performed and reported in a research study increase. An effect size (i.e., *d*) is also reported as recommended by the APA *Publication Manual*. This information is valuable to researchers doing meta-analyses and who wish to compare results of studies using similar variables. On the other hand, confidence intervals provide a range of possible effect sizes in terms of actual mean differences and not a single value such as Cohen's *d*. Because zero is not within the interval, we know that the outcome would be statistically significant at the .05 level (see Chapter 12). However, as the APA *Manual* emphasizes, confidence intervals provide information about precision of estimation and location of an effect that is not given by NHST alone. Recall from Chapter 12 that the smaller the confidence interval the more precise is our estimate.

Note: As mentioned in Chapter 12, a figure usually is not needed when only two group means are involved. Pictorial summaries such as graphs become more important when summarizing the results of experiments with more than two groups. If a figure were drawn showing the mean performance in the groups, then the statement of results should refer to the figure. (See Chapter 14.)

Power Analysis When we know the effect size, we can determine the statistical power of an analysis. Power, as you will recall, is the probability that a statistically significant effect will be obtained. Suppose that a previous study of vocabulary size contrasting younger and older adults produced an effect size of .50, a medium effect according to Cohen's (1988) rule of thumb. We can use the power tables created by Cohen to determine the number of participants needed in a test of mean differences to "see" an effect of size .50 with alpha .05. (See Appendix Table A.4.) The table identifies the power associated with various effect sizes as a function of sample size. Looking at the power table, we see that the sample size (in each group) of a two-group study would have to be about 64 to achieve power of .80 (for a two-tailed test). Looking for a medium effect size, we would need a total of 128 (64×2) participants to obtain statistical significance in 8 of 10 tries. Had the researchers been looking for a medium effect, their vocabulary study would have been underpowered. As it turns out, anticipating a large effect size, a sample size of 26 was appropriate to obtain power .80.

If the result is not statistically significant, then an estimate of power should be reported. If, for example, using an independent groups design the outcome had been $t(28) = 1.96$, $p > .05$, with an effect size of .50, we can determine the power of the study after the fact. Assuming equal-size groups in the study, we know that there were 15 subjects in each group ($df = n_1 + n_2 - 2$, or $28 = 15 + 15 - 2$). An examination of Appendix Table A.4 reveals that power for this study is .26. A statistically significant outcome would be obtained in only about 1 of 4 attempts with this sample size and when a medium (.50) effect

STRETCHING EXERCISE
A TEST OF (YOUR UNDERSTANDING OF) THE NULL HYPOTHESIS TEST

As should be apparent by now, understanding, applying, and interpreting results of NHST is no easy task. Even seasoned researchers occasionally make mistakes. To help you avoid mistakes, we provide a true-false test based on the information presented thus far about NHST.

Assume that an independent groups design was used to assess performance of participants in an experimental and control group. There were 12 participants in each condition, and results of NHST with alpha set at .05 revealed $t(22) = 4.52, p = .006$. True or false? The researcher may reasonably conclude on the basis of this outcome that

1. The null hypothesis should be rejected.
2. The research hypothesis has been shown to be true.
3. The results are of scientific importance.
4. The probability that the null hypothesis is true is only .006.
5. The probability of finding statistical significance at the .05 level if the study were replicated is greater than if the exact probability had been .02.

must be found. In this case, researchers would need to decide if practical or theoretical decisions should be made on the basis of this result or if "more research is needed."

Data Analysis Involving More Than Two Conditions

Thus far we have discussed the stages of data analysis in the context of an experiment with two conditions, that is, two levels of one independent variable. What happens when we have more than two levels (conditions) or, as is often the case in psychology, more than two independent variables? The most frequently used statistical procedure for analyzing results of psychology experiments in these situations is the analysis of variance (ANOVA).

We illustrate how ANOVA is used to test null hypotheses in four specific research situations: single-factor analysis of independent groups designs; single-factor analysis for repeated measures designs; two-factor analysis for independent groups designs; and two-factor analysis for mixed designs. We recommend that, before proceeding, you review the information presented in Chapters 7, 8, and 9 that describes these research designs.

ANOVA for Single-Factor Independent Groups Design

- Analysis of variance (ANOVA) is an inferential statistics test used to determine whether an independent variable has had a statistically significant effect on a dependent variable.
- The logic of analysis of variance is based on identifying sources of error variation and systematic variation in the data.
- The *F*-test is a statistic that represents the ratio of between-group variation to within-group variation in the data.
- The results of the initial overall analysis of an omnibus *F*-test are presented in an analysis of variance summary table; comparisons of two means can then be used to identify specific sources of systematic variation in an experiment.
- Although analysis of variance can be used to decide whether an independent variable has had a statistically significant effect, researchers examine the descriptive statistics to interpret the meaning of the experiment's outcome.
- Effect size measures for independent groups designs include eta squared (η^2) and Cohen's *f*.
- A power analysis for independent groups designs should be conducted prior to implementing the study in order to determine the probability of finding a statistically significant effect, and power should be reported whenever nonsignificant results based on NHST are found.
- Comparisons of two means may be carried out to identify specific sources of systematic variation contributing to a statistically significant omnibus *F*-test.

Key Concept

Overview Statistical inference requires a test to determine whether or not the outcome of an experiment was statistically significant. The most commonly used inferential statistics test in the analysis of psychology experiments is the **ANOVA.** As its name implies, the analysis of variance is based on analyzing different sources of variation in an experiment. In this section we briefly

Key Concept

introduce how the analysis of variance is used to analyze experiments that involve independent groups with one independent variable, or what is called a **single-factor independent groups design.** Although ANOVA is used to analyze the results of either random groups or natural groups designs, the assumptions underlying ANOVA strictly apply only to the random groups design.

There are two sources of variation in any random groups experiment. First, variation within each group can be expected because of individual differences among subjects who have been randomly assigned to a group. The variation due to individual differences cannot be eliminated, but this variation is presumed to be balanced across groups when random assignment is used. In a properly conducted experiment, the differences among subjects within each group should be the only source of error variation. Participants in each group should be given instructions in the same way, and the level of the independent variable to which they've been assigned should be implemented in the same way for each member of the group (see Chapter 7).

The second source of variation in the random groups design is variation between the groups. If the null hypothesis is true (no differences among groups), any observed differences among the means of the groups can be attributed to error variation (e.g., the different characteristics of the participants in the groups). As we've seen previously, however, we don't expect sample means to be exactly identical. Fluctuations produced by sampling error make it likely that the means will vary somewhat—this is error variation. Thus, the variation among the different group means, when the null hypothesis is assumed to be true, provides a second estimate of error variation in an experiment. If the null hypothesis is true, this estimate of error variation *between* groups should be similar to the estimate of error variation *within* groups. Thus, the random groups design provides two independent estimates of error variation, one within the groups and one between the groups.

Now suppose that the null hypothesis is false. That is, suppose the independent variable has had an effect in your experiment. If the independent variable has had an effect, should the means for the different groups be the same or different? You should recognize that they should be different. An independent variable that has an effect on behavior should produce systematic differences in the means across the different groups of the experiment. That is, the independent variable should introduce a source of variation among the groups of the experiment—it should cause the groups to vary. This systematic variation will be added to the differences in the group means that are already present due to error variation. That is, between-group variation will increase.

The *F*-Test We are now in a position to develop a statistic that will allow us to tell whether the variation due to our independent variable is larger than would be expected on the basis of error variation alone. This statistic is called *F*; it is named after Ronald Fisher, the statistician who developed the test. The conceptual definition of the ***F*-test** is

Key Concept

$$F = \frac{\text{Variation between groups}}{\text{Variation within groups}} = \frac{\text{Error variation + systematic variation}}{\text{Error variation}}$$

If the null hypothesis is true, there is no systematic variation between groups (no effect of the independent variable) and the resulting *F*-test has an expected value of 1.00 (since error variation divided by error variation would equal 1.00). As the amount of systematic variation increases, however, the expected value from the *F*-test becomes greater than 1.00.

The analysis of experiments would be easier if we could isolate the systematic variation produced by the independent variable. Unfortunately, the systematic variation between groups comes in a "package" along with error variation. Consequently, the value of the *F*-test may sometimes be larger than 1.00 simply because our estimate of error variation between groups happens to be larger than our estimate of error variation within groups (i.e., the two estimates should be similar but can differ due to chance factors). How much greater than 1.00 does the *F* statistic have to be before we can be relatively sure that it reflects true systematic variation due to the independent variable? Our earlier discussion of statistical significance provides an answer to this question. To be statistically significant, the *F* value needs to be large enough so that its probability of occurring if the null hypothesis were true is less than our chosen level of significance, usually .05.

We are now ready to apply the principles of NHST and the procedures of ANOVA to analyze a specific experiment.

Analysis of Single-Factor Independent Groups Design The first step in doing an inferential statistics test like the *F*-test is to state the research question the analysis is intended to answer. Typically, this takes the form of "Did the independent variable have any overall effect on performance?" Once the research question is clear, the next step is to develop a null hypothesis for the analysis. The experiment we will discuss as an example examines the effect on memory retention of several kinds of memory training. There are four levels (conditions) of this independent variable and, consequently, four groups of participants. Each sample or group represents a population. The initial overall analysis of the experiment is called an **omnibus *F*-test.** The null hypothesis for such omnibus tests is that all the population means are equal. Remember that the null hypothesis assumes no effect of the independent variable. The formal statement of a null hypothesis (H_0) is always made in terms of population characteristics. These population characteristics are indicated by Greek letters, and the population mean is symbolized as μ ("mu"). We can use a subscript for each mean to represent the levels of the independent variable. Our null hypothesis then becomes

Key Concept

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$$

The alternative to the null hypothesis is that one or more of the means of the populations are not equal. In other words, the alternative hypothesis (H_1) states that H_0 is wrong; there is a difference somewhere. The alternative hypothesis becomes

$$H_1: \text{NOT } H_0$$

TABLE 13.2 NUMBER OF WORDS RECALLED IN A MEMORY EXPERIMENT

	Instruction (A)							
Subject	Control (a_1)	Subject	Story (a_2)	Subject	Imagery (a_3)	Subject	Rhymes (a_4)
1	12	6	15	11	16	16	14
2	10	7	14	12	16	17	14
3	9	8	13	13	13	18	15
4	11	9	12	14	12	19	12
5	8	10	12	15	15	20	12
Mean	10.0		13.2		14.4		13.4
Standard deviation	1.6		1.3		1.8		1.3
Range	8–12		12–15		12–16		12–15

If the type of memory training does have an effect on retention (i.e., if the independent variable produces systematic variation), then we will want to reject the null hypothesis.

The data in Table 13.2 represent the number of words correctly recalled (out of a possible 20) on a retention test in an experiment investigating memory training techniques. Five participants were randomly assigned to each of four groups (defined by the method of study that individuals were instructed to use to learn the words in preparation for the memory test). The control method involved no specific instructions, but in the three experimental groups participants were instructed to study by making up a story using the to-be-remembered words (story method), to use visual imagery (imagery method), or to use rhymes to remember the words (rhyme method). The independent variable being manipulated is "instruction," and it can be symbolized by the letter "A." The levels of this independent variable can be differentiated by using the symbols a_1, a_2, a_3, and a_4 for the four respective groups. The number of participants within each group is referred to as n; in this case, $n = 5$. The total number of individuals in the experiment is symbolized as N; in this case, $N = 20$.

An important step in the analysis of any experiment is to set up a data matrix like the one in Table 13.2. The number of correct responses is listed for each person in each of the four groups with each participant identified with a unique subject number. In order to understand the results of an experiment, it is essential to summarize the data prior to examining the outcome of the ANOVA. Below the data matrix the mean, range (minimum and maximum scores), and standard deviation are provided for each group.

Before examining the "significance" of any inferential test, try to get an impression of what the summary statistics are telling you. Look to see if there is a visible "effect" of the independent variable; that is, see if there is substantial variation among the means. By examining the ranges and standard deviations, get a sense of the variability in each group. (Remember, the less scores vary around their sample means, the better the chance of seeing an effect that is present.) The range, or difference between the minimum and maximum values,

TABLE 13.3 ANALYSIS OF VARIANCE SUMMARY TABLE FOR MEMORY EXPERIMENT

Source	Sum of squares	*df*	Mean square	*F*-ratio	*p*
Group	54.55	3	18.18	7.80	0.002
Error	37.20	16	2.33		
Total	91.75	19			

is useful in identifying floor and ceiling effects. Is the variability among the groups similar? We want the variation to be relatively homogeneous as wide discrepancies in within-group variability can create interpretation problems when using ANOVA.

Our examination of the summary statistics reveals that there appears to be systematic variation among the means; the largest difference is seen between the Control (10.0) and the Imagery Group (14.4). All the experimental means are larger than the Control mean. Note that the range is similar for all the groups; the standard deviations, too, are fairly similar. This attests to the homogeneity (similarity) of variance among the groups. (Many computer programs provide a test of "homogeneity of variance" along with the ANOVA output.) Moreover, an inspection of the highest scores in each group shows that ceiling effects are not a problem in this data set (as total possible was 20).

The next step in an analysis of variance is to do the computations to obtain the estimates of variation that make up the numerator and denominator of the *F*-test. Calculations for *F*-tests are best done using a computer. We will focus, therefore, on interpreting the results of the computations. The results of an analysis of variance are presented in *Analysis of Variance Summary Table* (see Table 13.3).

Interpreting the ANOVA Summary Table The summary table for the omnibus *F*-test for the independent groups design used to investigate the effect of memory training is found in Table 13.3. Remember that there were four groups of size $n = 5$ and, thus, overall $N = 20$. It is critically important you know what the ANOVA summary table contains. Thus, we examine the components of the summary table before looking at the outcome of the *F*-test for the experiment.

The left column of the summary table lists the two sources of variation described earlier. In this case the independent variable of the training group ("Group") is a source of variation between the groups, and the within-groups differences provide an estimate of error variation. The total variation in the experiment is the sum of the variation between and within groups. The third column is the degrees of freedom (*df*). In general, the statistical concept of degrees of freedom is defined as the number of entries of interest minus 1. Since there are 4 levels of the training independent variable, there are 3 *df* between groups. There are 5 participants within each group, so there are 4 *df* $(n - 1)$ within each of the 4 groups. Because all 4 groups are the same size, we can determine the within-groups *df* by multiplying the *df* within each group by the number of groups (4×4) for 16 *df*. The total *df* is the number of subjects minus 1 $(N - 1)$, or the sum of *df* between groups plus *df* within groups $(3 + 16 = 19)$.

The sums of squares (*SS*) and the mean square (*MS*) are computational steps in obtaining the *F* statistic. The *MS* between groups (row 1) is an estimate of systematic variation plus error variation and is calculated by dividing the *SS* between groups by the *df* between groups (54.55/3 = 18.18). The *MS* within groups (row 2) is an estimate of error variation only and is computed by dividing the *SS* within groups by the *df* within groups (37.20/16 = 2.33). The *F*-test is calculated by dividing the *MS* between groups by the *MS* within groups (18.18/2.33 = 7.80).

We are now ready to use the information in the summary table to test for the statistical significance of the outcome in the memory training experiment. You may anticipate the conclusion already, knowing that when the null hypothesis is assumed to be true (i.e., no effect of the independent variable), the estimate of systematic variation plus error variation (numerator of the *F*-test) should be approximately equal to the estimate of error variation only (denominator of the *F*-test). As we see here, the estimate of systematic variation plus error variation (18.18) is quite a bit larger than the estimate of error variation alone (2.33).

The obtained *F* value in this analysis (7.80) appears in the second to last column of the summary table. The probability of obtaining an *F* as large as 7.80 if the null hypothesis were true is shown in the last column of the summary table (0.002). The obtained probability of .002 is less than the level of significance ($\alpha = .05$), so we reject the null hypothesis and conclude that the overall effect of memory training is statistically significant. The results of NHST using ANOVA would be summarized in your research report as

$$F(3, 16) = 7.80, p = .002$$

An *F* statistic is identified by its degrees of freedom. In this case there are 3 *df* between groups and 16 *df* within groups (i.e., 3, 16). Note that the exact probability (i.e., .002) is reported because it gives us information about the probability of replication.

Just what have we learned when we find a statistically significant outcome in an analysis of variance testing an omnibus null hypothesis? In one sense, we have learned something very important. We are now in a position to state that manipulating the independent variable produced a change in performance (i.e., participants' memory for the to-be-remembered words). In another sense, merely knowing our outcome is statistically significant tells us little about the nature of the effect of the independent variable. The descriptive statistics (in our example, the mean number of words recalled as reported in Table 13.2) allow us to describe the nature of the effect. Note that only by examining the pattern of group means do we begin to learn what happened in our experiment as a function of the independent variable. *Never try to interpret a statistically significant outcome without referring to the corresponding descriptive statistics*.

Although we know that the omnibus *F*-test was statistically significant, we do not know the degree of relationship between the independent and dependent variables, and thus we should consider calculating an effect size for our independent variable. Based on the omnibus test alone we also are unable to state which of the group means differed significantly. Fortunately, there are analysis techniques that allow us to locate more specifically the sources of

systematic variation in our experiments. One approach that is highly recommended is the use of confidence intervals (see Chapter 12). Confidence intervals can provide evidence for the pattern of population means estimated by our samples (see especially Box 12.5). Another technique is that of comparing two means. We first discuss an effect size measure for the independent groups ANOVA, as well as power analysis for this design, and then turn our attention to comparisons of two means.

Calculating Effect Size for Designs with Three or More Independent Groups

Key Concept

We mentioned earlier that the psychology literature contains many different measures of effect magnitude, which depend on the particular research design, test statistic, and other peculiarities of the research situation (e.g., Cohen, 1992; Kirk, 1996; Rosenthal & Rosnow, 1991). When we know one measure of effect magnitude, we usually can translate it to another, comparable measure without much difficulty. An important class of effect magnitude measures that applies to experiments with more than two groups is based on measures of "strength of association" (Kirk, 1996). What these measures have in common is that they allow estimates of the proportion of total variance accounted for by the effect of the independent variable on the dependent variable. A popular strength of association measure is **eta squared,** or η^2. It is easily calculated based on information found in the ANOVA Summary Table (Table 13.3) for the omnibus *F*-test (although many computer programs automatically provide eta squared as a measure of effect size). Eta squared is defined as

$$\frac{\text{Sum of squares between groups}}{\text{Total sum of squares}}$$

In our example (see Table 13.3),

$$\text{eta squared } (\eta^2) = \frac{54.55}{[(54.55) + (37.20)]} = .59$$

Eta squared can also be computed directly from the *F*-ratio for the between-groups effect when the ANOVA table is not available (see Rosenthal & Rosnow, 1991, p. 441):

$$\text{eta squared } (\eta^2) = \frac{(F)(df \text{ effect})}{[(F)(df \text{ effect})] + (df \text{ error})}$$

or, in our example,

$$\text{eta squared } (\eta^2) = \frac{(7.80)(3)}{[(7.80)(3)] + 16} = .59$$

Another measure, designed by J. Cohen, for designs with three or more independent groups is f (see Cohen, 1988). It is a standardized measure of effect size similar to d, which we saw was useful for assessing effect sizes in a two-group experiment. However, unlike d, which defines an effect in terms of

Key Concept

the difference between two means, **Cohen's *f*** defines an effect in terms of a measure of dispersal among group means. Both *d* and *f* express the effect relative to (i.e., "standardized" on) the within-population standard deviation. Cohen has provided guidelines for interpreting *f*. Specifically, he suggests that small, medium, and large effects sizes correspond to *f* values of .10, .25, and .40. The calculation of *f* is not easily accomplished using the information found in the ANOVA Summary Table (Table 13.3), but it can be obtained without much difficulty once eta squared is known (see Cohen, 1988), as

$$f = \sqrt{\frac{\eta^2}{1 - \eta^2}}$$

or, in our example,

$$f = \sqrt{\frac{.59}{1 - .59}} = 1.20$$

We can thus conclude that memory training accounted for .59 of the total variance in the dependent variable and produced a standardized effect size, *f*, of 1.20. Based on Cohen's guidelines for interpreting *f* (.10, .25, .40), it is apparent that memory training had a large effect on recall scores.

Assessing Power for Independent Groups Designs

Once the effect size is known, we can obtain an estimate of power for a specific sample size and degrees of freedom associated with the numerator (between-groups effect) of the *F*-ratio. In our example, we set alpha at .05; the experiment was done with $n = 5$ and $df = 3$ for the between-groups effect (number of groups minus 1). The effect size, *f*, associated with our data set is very large (1.20), and there is no good reason to conduct a power analysis for this large effect which was statistically significant. Thus, let's consider a somewhat different outcome.

Assume that the ANOVA in our example yielded a nonsignificant *F* and effect size was $f = .40$, still a large effect according to Cohen's guidelines. An important question to answer is "What was the power of our experiment?" How likely were we to see an effect of this size given an alpha of .05, a sample size of $n = 5$, and $df = 3$ for our effect? Consulting the abbreviated power table in the Appendix (see Table A.5), we find that under these conditions power was .26. In other words, the probability of obtaining statistical significance in this situation was only .26. In only approximately one fourth of the attempts under these conditions would we obtain a significant result. The experiment was clearly underpowered, and it is unreasonable to make much of the fact that NHST did not reveal a significant result. To do so would ignore the very important fact that the effect of our independent variable was, in fact, large.

Although learning about power after the fact can be important, particularly when we obtain a nonsignificant outcome based on NHST, ideally power analysis should be conducted prior to an experiment in order to reveal the a priori (from the beginning) probability of finding a statistically significant effect. An

experimenter who begins an experiment knowing that power is only .26 would appear to be wasting time and resources given that the odds of *not* finding a significant effect are .74. Let us assume, therefore, that the experiment has not yet been conducted and that the investigator examined the literature on memory training and found that a large effect was often obtained by previous researchers in this area. Let us further assume that the researcher wants power to be .80 in the experiment. Because power is typically increased by increasing sample size, the researcher will want to find out what the sample size should be in order to find a large effect with power .80.

To find the sample size needed to see a large effect in 8 of 10 tries, we use the power table to find n with $df = 3$ under the effect size heading of .40. Appendix Table A.5 shows that to achieve power of .80 we would need about 18 participants in each condition of the experiment. The researcher should take this information into consideration before doing the experiment.

Comparing Means in Multiple-Group Experiments

As we noted, knowing that "something happened" in a one-factor, multiple-group experiment is often not very interesting. We generally do research, or at least we should, with more specific hypotheses in mind than "this variable will have an effect" on the dependent variable. Neither the results of the omnibus F, nor a measure of overall effect size, tell us which means are significantly different from which other means. We cannot, for instance, look at the four means in our memory experiment and judge that the "imagery" mean is significantly different from the "story" mean. The results of the omnibus F simply tell us there is variation present among all the groups that is larger than would be expected by chance in this situation.

We can suggest two complementary ways to learn more about what happened in a multiple-group, single-factor experiment. One approach is to examine the probable pattern of population means by calculating 95% confidence intervals for the mean estimates in our experiment. This approach was illustrated in Chapter 12 when we showed how confidence intervals could be used to compare means in a multiple-group experiment. Confidence intervals can be used to make decisions about the probable differences among population means that are estimated by the means of our experimental groups. These decisions are made by examining whether confidence intervals overlap, and if they do, to what degree they overlap (see especially Box 12.5). Remember that the width of the confidence intervals provides information about the precision of our estimates.

Key Concept

A second approach makes use of NHST and focuses on a small set of two-group comparisons in order to specify the source of the overall effect of our independent variable. A **comparison of two means** allows the researcher to focus on a particular difference of interest. These comparisons can be quite sophisticated, for example, comparing the average of two or more groups in an experiment with the mean of another group or the average of two or more other groups. However, most of the time we will be interested in the difference between just two means that are represented by individual groups. These two-mean comparisons are usually made after we have determined that our omnibus F-test is statistically significant.

One approach for carrying out comparisons of two means is to use a *t*-test; however, there is a slight modification in the way that *t* is calculated when comparing means in a multiple-group experiment. Specifically, we want to use a *pooled variance* estimate based on the within-group variation estimate (MS_{error}) found in our omnibus *F*-test. That is, our variance estimate uses information obtained from *all* the groups in our experiment, not just the two groups of interest. Therefore, the formula for this *t*-test is

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{[MS_{error}]\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}}$$

The value for the MS_{error} is obtained from the ANOVA Summary Table of our omnibus *F*-test, and the degrees of freedom for our comparison *t*-test are those associated with the MS_{error} [or $k(n - 1)$, where k = number of groups]. For example, the *MS* within groups (error) for the analysis reported in Table 13.3 is 2.33 with 16 degrees of freedom [$4(5 - 1) = 16$].

One comparison of two means we could make for the memory experiment is to compare the mean performance for the memory-training groups (combined) and the control group. The mean retention for the three memory training groups is 13.67 ($n = 15$), and the mean for the control group is 10.00 ($n = 5$). We can ask, does memory training, regardless of type (i.e., story, imagery, rhymes), lead to better memory retention than no memory training (control)? The null hypothesis is that the two population means do not differ (and the sample means differ by chance alone). When the appropriate values are substituted into the formula for *t* given above, we observe a statistically significant effect, $t(16) = 4.66$, $p = .0003$.[2] Thus, memory training in this experiment, regardless of type, resulted in better memory retention for the words compared to no training. You can see that this statement is more specific than the statement we could make based on the omnibus *F* test, in which we could say only that the variation across the four conditions of the experiment was larger than that expected based on chance alone. Our discussion of the confidence intervals for these means in the next section will help you to determine whether it would be useful to conduct additional two-mean comparisons among the three memory-training conditions.

Cohen's *d* may be calculated for comparisons of two means using the results of the *t*-test. The formula for Cohen's *d* in this situation is

$$d = \frac{2(t)}{\sqrt{df_{error}}}$$

[2]Often the *t*-tests for comparisons of two means following an omnibus *F*-test can be easily calculated by hand using the formula provided above. The probability (*p*) associated with the computed value of *t* can then be determined in one of two ways: (1) compare the observed *t* to the critical values of *t* found in Appendix Table A.2, or (2) use a computer program that allows users to enter the *df* and *t* values for a one- or two-tailed test to obtain the exact probability associated with those values (e.g., try the website at www.danielsoper.com/statcalc/calc08.aspx).

For the comparison between the three memory-training groups and the control group, substituting the value of 4.66 into the formula and with 16 df_{error}, the effect size, d, is 2.33. According to Cohen's criteria for effect sizes, this can be interpreted as a large effect of memory instruction relative to no instruction.

We can conclude this section on analyzing mean differences in a multiple-group experiment by reviewing the complementary information obtained by using confidence intervals and NHST. Because each approach adds unique kinds of information, an argument can be made for using both confidence intervals and NHST in this situation. The use of confidence intervals allows us to make decisions about the probable pattern of population means across all the conditions of our experiment. The width of the interval tells us how precisely we have estimated the population mean. When using a *t*-test we are seeking to make a decision about rejecting or not rejecting the null hypothesis with a specific probability (e.g., $p = .001$). As noted previously, the exact probability associated with the outcome of NHST can be important when interpreting results (e.g., Posavac, 2002). The lower the exact probability, the greater is the likelihood that an exact replication would permit rejecting the null hypothesis at $p < .05$ (see Zechmeister & Posavac, 2003). Minimally, we want to report the lowest probability for statistical significance for which we have information. (Computers automatically give the exact probability of our test result.)

The results of the *t* comparison also permit us to contrast results with previous studies in two ways. First, we can note whether our experiment's findings for statistical significance are similar to those observed in a previous experiment. That is, did we *replicate* a statistically significant finding? Second, we can calculate an effect size (e.g., Cohen's d) for this two-mean comparison that may be compared with effects obtained in previous experiments, perhaps as part of a meta-analysis. Neither of these contrasts is easy to do using confidence intervals. That is, unlike NHST, confidence intervals do not provide an exact probability associated with a difference seen in our experiment and the calculation of an effect size is more directly carried out following a *t*-test (see Chapter 12).

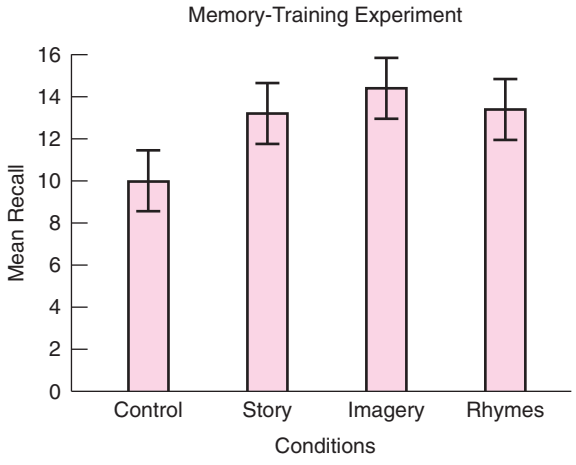
In summary, we encourage you to look at your data and differences between means using more than one statistical technique, seeking evidence for "what happened" from different approaches to data analysis.

Reporting Results of a Single-Factor Independent Groups Experiment

We are now ready to model a statement of results based on the recommendations of the APA *Publication Manual* (2001; see especially pp. 20–26). The report includes information obtained from both the construction of confidence intervals (Chapter 12) and NHST. A report of the results of the memory training study might take this form:

> The effect of memory training on retention of words was examined. Mean recall (out of a possible 20 words) was determined for each of the four instructional conditions, each with five participants ($N = 20$): story method, imagery method, rhyme method, and control (no specific instructions). These means (and sample standard deviations) were, for each of the conditions: story, 13.2 (1.3); imagery,

FIGURE 13.1 Means and 95% confidence intervals for the memory-training experiment.



14.4 (1.8); rhyme, 13.4 (1.3); control, 10.0 (1.6). Mean recall differed significantly among the four instruction conditions, $F(3, 16) = 7.80$, $p = .002$, $\eta^2 = .59$. A comparison of two means was performed to contrast the overall mean recall of the three memory training groups with that of the control group. Retention of words was greater for participants in the memory training groups (13.7) compared to the control group, $t(16) = 4.66$, $p < .01$. The effect size, d, for this comparison was 2.33, indicating a large effect of memory training relative to no training. The 95% confidence interval for the mean of each group is shown in Figure 13.1. The confidence interval for the control group does not overlap the intervals for the instructional groups. However, the intervals among the three training conditions overlap substantially (including the sample means within these intervals), indicating the population means for the three instructional conditions are not likely to differ. This pattern indicates that instructions to use specific memory techniques, regardless of type of technique, were successful in increasing memory retention relative to a noninstructed control group.

Commentary In the first sentence of the report we find information about the purpose of the experiment, the overall number of participants (20), the number of levels of the independent variable, how levels were defined (and their names), and the size of each group (5). Descriptive statistics are then provided for each group. The width of the confidence intervals calls our attention to the precision of estimation (or lack of) of the population means for each group as well as the likely pattern of population means. The construction of confidence intervals follows the procedure outlined in Chapter 12. Because the square root of the MS_{error} from the ANOVA summary table is equivalent to s_{pooled}, we can define the 95% confidence interval as

$$95\%\ CI = \overline{X} \pm [\sqrt{(MS_{error}/n)}](t_{crit})$$

where t_{crit} is the value for t with degrees of freedom associated with the MS_{error}.

In our example, the degrees of freedom for MS_{error} are 16 (see ANOVA Summary Table) and t_{crit} at the .05 level (two-tailed test) is 2.12. Therefore,

$$95\%\ CI = \overline{X} \pm [\sqrt{(2.32/5)}](2.12) = \overline{X} \pm \left(\sqrt{.464}\right)(2.12) = \overline{X} \pm (.68)(2.12)$$
$$= \overline{X} \pm 1.44$$

As shown in Figure 13.1, intervals overlap among the three training conditions but do not overlap with the control interval. The report of the omnibus *F* is accompanied by both the exact probability for the *F*-test and an effect size measure, eta squared. Neither of these pieces of information can be obtained from an examination of the confidence intervals. This information is repeated for the comparison of two means, which focuses on the comparison between the performance of the experimental groups and the control group. Usually only one effect size measure is reported and, as you saw, we chose *d*, although one might reasonably prefer eta squared. Finally, the final sentence summarizes the results of the experiment for the reader.

It may be useful in some situations to perform additional comparisons of two means, for example, contrasting the difference between one or more of the experimental groups and the control group. However, in this case, an examination of the confidence intervals in Figure 13.1 suggests that, since intervals overlap substantially among the three training conditions but none overlaps with the control interval, we may reasonably conclude that each specific training procedure differed from the control and the training conditions did not differ from one another (see Chapter 12).

Repeated Measures Analysis of Variance

- The general procedures and logic for null hypothesis testing using repeated measures analysis of variance are similar to those used for independent groups analysis of variance.
- Before beginning the analysis of variance for a complete repeated measures design, a summary score (e.g., mean, median) for each participant must be computed for each condition.
- Descriptive data are calculated to summarize performance for each condition of the independent variable across all participants.
- The primary way that analysis of variance for repeated measures differs is in the estimation of error variation, or residual variation; residual variation is the variation that remains when systematic variation due to the independent variable and subjects is removed from the estimate of total variation.

The analysis of experiments using repeated measures designs involves the same general procedures used in the analysis of independent groups design experiments. The principles of NHST are applied to determine whether the differences obtained in the experiment are larger than would be expected on the basis of error variation alone. The analysis begins with an omnibus analysis of variance to determine whether the independent variable has produced any

systematic variation among the levels of the independent variable. Should this omnibus analysis prove statistically significant, confidence intervals and comparisons of two means can be made to find the specific source of the systematic variation—that is, to determine which specific levels differed from each other. We have already described the logic and procedures for this general analysis plan for experiments that involve independent groups designs. We will focus in this section on the analysis procedures specific to repeated measures designs. Our example will be the time-perception experiment described in Chapter 8.

Summarizing the Data Recall that in a repeated measures design, each participant experiences every condition of the experiment. In a complete design, each participant experiences every condition more than once; in an incomplete design, each participant experiences every condition exactly once. In Chapter 8 we described an experiment in which participants estimated the duration of four time intervals (12, 24, 36, and 48 seconds) in a complete repeated measures design. For example, on a single trial, participants experienced a randomly determined time interval (e.g., 36 seconds) and then were asked to estimate the duration of the interval.

The first step for analyzing these data is to prepare a data matrix that allows you to summarize participants' performance in each condition of the experiment. In a complete design, this requires that, for each participant, you first calculate a score to summarize each individual's performance in each condition. This summary statistic balances the practice effects associated with any one particular trial across the trials for a condition. In the time-perception experiment, participants experienced each condition six times; thus, with four conditions in the experiment, each participant made 24 estimates. A median was used to summarize each participant's performance in each of the four conditions. Typically a mean would be calculated to summarize a participant's score within a condition, but recall that a mean is influenced by extreme scores. Participants' time estimates are likely to have extreme values on some trials (e.g., due to inattention). The median scores for five participants are presented in the top portion of Table 13.4. For example, the median amount of time estimated by the first participant for the 12-second interval condition, across six trials, was 13 seconds. The next step in summarizing the data is to calculate descriptive statistics across the participants for each of the conditions. The means and standard deviations (in parentheses) for each condition also appear in Table 13.4.

The focus of the analysis was on whether the participants could discriminate intervals of different lengths. As you probably have already realized, we cannot confirm the participants' ability to discriminate intervals of varying lengths until we know that the mean differences in Table 13.4 are greater than would be expected on the basis of error variation alone. That is, even though it may appear that participants were able to discriminate between the different intervals when examining the means, we do not know if their performance was different from that which would occur by chance. The null hypothesis for an omnibus analysis of variance for the data in Table 13.4 is that the population means estimated for each interval are the same. To perform an *F*-test of this null hypothesis, we need an estimate of error variation plus systematic variation (the numerator of an

TABLE 13.4 DATA MATRIX AND ANALYSIS OF VARIANCE SUMMARY TABLE FOR A REPEATED MEASURES DESIGN EXPERIMENT

Data matrix

	Interval length			
Participant	**12**	**24**	**36**	**48**
1	13	21	30	38
2	10	15	38	35
3	12	23	31	32
4	12	15	22	32
5	16	36	69	60
Mean (*SD*)	12.6 (2.0)	22.0 (7.7)	38.0 (16.3)	39.4 (10.5)

Note: Each value in the table represents the median of the participants' six responses at each level of the interval-length variable.

Source of variation	*df*	*SS*	*MS*	*F*	*p*
Subjects	4	1553.5	—	—	—
Interval length	3	2515.6	838.5	15.6	.000
Residual (error variation)	12	646.9	53.9		
Total	19	4716.0			

F-test). The variation among the mean estimates across participants for the four intervals provides the information we need for the numerator. Even if these five participants had been tested on only one interval length several times, we would not expect their mean estimates to be identical. Thus, we know the mean estimates we have for each level of the interval variable reflect error variation as well. We also know, however, that if the different interval lengths did systematically affect the participants' judgments, then the mean estimates for the intervals would reflect this systematic variation. To complete the *F*-test, we also need an estimate of error variation alone (the denominator of the *F*-test). See Box 13.4.

The source of error variation in the repeated measures designs is the differences in the ways the conditions affect different participants. Perhaps the best way to describe the way we get these estimates is to say that we do it "by default." We first determine how much total variation there is in our experiment. Then we subtract the two potential sources of systematic variation: the independent variable and subjects. The remainder is called residual variation, and it represents our estimate of error variation alone. As was the case in the random groups design when we used variation within groups as our estimate of error variation alone, residual variation serves as the denominator for the *F*-test in repeated measures designs (i.e., as an estimate of error variation alone).

Interpreting the ANOVA Summary Table The analysis of variance summary table for this analysis is presented in the lower portion of Table 13.4. The computations of a repeated measures analysis of variance would be done using a statistical software package on a computer. Our focus now is on interpreting the values in the

BOX 13.4

ESTIMATING ERROR AND SENSITIVITY IN A REPEATED MEASURES DESIGN

One distinctive characteristic of the analysis of repeated measures designs is the way in which error variation is estimated. We described earlier that for the random groups design individual differences among participants that are balanced across groups provide the estimate of error variation that becomes the denominator of the *F*-test. Because individuals participate in only one condition in these designs, differences among participants cannot be eliminated—they can only be balanced. In repeated measures designs, on the other hand, there is systematic variation among participants. Some participants consistently perform better across conditions, and some participants consistently perform worse. Because each individual participates in each condition of repeated measures designs, however, differences among participants contribute equally to the mean performance in each condition. Accordingly, any differences among the means for each condition in repeated measures designs cannot be the result of systematic differences among participants. In repeated measures designs, however, differences among participants are not just balanced—they are actually eliminated from the analysis. *The ability to eliminate systematic variation due to participants in repeated measures designs makes these designs generally more sensitive than random groups designs.*

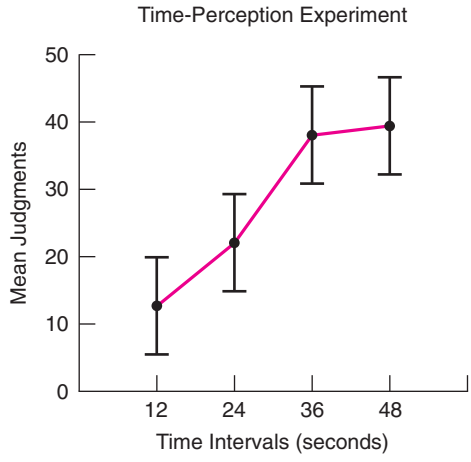
summary table and not on how these values are computed. Table 13.4 lists the four sources of variation in the analysis of a repeated measures design with one manipulated independent variable. Reading from the bottom of the summary table up, these sources are (1) total variation, (2) residual variation, (3) variation due to interval length (the independent variable), and (4) variation due to subjects.

As in any summary table, the most critical pieces of information are the *F*-test for the effect of the independent variable of interest and the probability associated with that *F*-test assuming the null hypothesis is true. The important *F*-test in Table 13.4 is the one for interval length. The numerator for this *F*-test is the mean square (*MS*) for interval length; the denominator is the residual *MS*. There are four interval lengths, so there are 3 degrees of freedom (*df*) for the numerator. There are 12 *df* for the residual variation. We can obtain the *df* for the residual variation by subtracting the *df* for subjects and for interval length from the total *df*($19 - 4 - 3 = 12$). The obtained *F* of 15.6 has a probability under the null hypothesis of .000, which is less than the .05 level of significance we have chosen as our criterion for statistical significance. So we reject the null hypothesis and conclude that the interval length was a source of systematic variation. This means that we can conclude that the participants' estimates did differ systematically as a function of interval length.

Figure 13.2 shows 95% confidence intervals around the means in the time-perception experiment. The procedure for constructing these intervals is the same as that for the independent groups experiment. Intervals were constructed using the MS_{error} (residual) in the omnibus ANOVA (as recommended by Loftus & Masson, 1994). That is,

$$95\%\ CI = \overline{X} \pm \left[\sqrt{(MS_{error}/n)}\right](t_{crit})$$

FIGURE 13.2 Means and 95% confidence intervals for the time-perception experiment.



where t_{crit} is the value of t with the degrees of freedom associated with the MS_{error} (residual). The interpretation of confidence intervals in the repeated measures design is the same as that of the independent groups design (see Chapter 12).

Effect Size Measures As we mentioned previously, it is a good idea to include measures of effect size for your analyses. A typical measure of effect size for a repeated measures design is the strength of association measure called eta squared (η^2). It may be calculated by dividing the sum of squares for the within-subjects effect by the combined sums of squares for the within-subjects effect and residual or error. For our sample study,

$$\text{eta squared } (\eta^2) = \frac{SS_{effect}}{SS_{effect} + SS_{error}} = \frac{2515.6}{2515.6 + 646.9} = .795$$

This indicates the proportion of variance accounted for by the independent variable.

In some cases, the omnibus analysis of variance would be followed by comparisons of two means (such as comparing the mean for each interval to the mean for the succeeding interval) to determine more exactly that mean estimates increased with increasing interval lengths. Once again, the logic of these two-mean comparisons corresponds to the logic we considered for comparisons in the random groups design. The decision to perform these comparisons will be influenced by the specific hypotheses being tested in the experiment and by knowledge gained from an examination of confidence intervals surrounding the means. (See Keppel [1991], however, for a discussion of the complications that can arise in doing two-mean comparisons in repeated measures designs.)

Two-Factor Analysis of Variance for Independent Groups Designs

The two-factor analysis of variance for independent groups designs is used for the analysis of experiments in which each of two independent variables was manipulated at two or more levels. The logic of complex designs with two independent variables and the conceptual basis for the analysis of these experiments are described in Chapter 9. In Chapter 9 you also learned to describe both main effects and interaction effects. We will focus in this chapter on the computer-assisted analysis of a factorial design that involves *F*-tests for the main effect of A, the main effect of B, and the interaction effect, A × B. The two-factor analysis for independent groups is applicable to experiments in which both independent variables are manipulated using a random groups design, in which both independent variables represent the natural groups design, or in which one independent variable represents the natural groups design and the other represents the random groups design. The analysis of a complex design proceeds somewhat differently depending on whether the omnibus *F*-test does or does not reveal an interaction effect. We first consider the analysis plan when an interaction effect is detected.

Analysis of a Complex Design with an Interaction Effect

- If the omnibus analysis of variance reveals a statistically significant interaction effect, the source of the interaction effect is identified using simple main effects analyses and comparisons of two means.
- A simple main effect is the effect of one independent variable at one level of a second independent variable.
- If an independent variable has three or more levels, comparisons of two means can be used to examine the source of a simple main effect by comparing means two at a time.
- Confidence intervals may be drawn around group means to provide information regarding the precision of estimation of population means.

Rodman and Burger (1985) investigated a phenomenon called the defensive attribution effect. Attributions refer to the causal explanations we use to try to account for our own behavior and the behavior of others. In previous experiments, nondepressed participants who read a description of an accident in which a person suffered mild or severe consequences attributed more responsibility to the perpetrator in the severe than in the mild accident condition. It appears that nondepressed individuals do not want to attribute the cause of a severe accident to chance. To do so would increase the perceived possibility that they themselves could be involved in a severe accident. The attribution of greater responsibility to the perpetrator in the severe accident is called defensive attribution. Nondepressed individuals "defended" or protected themselves against thinking that they would be involved in future severe accidents. Because depressed individuals tend to exhibit negative thinking, Rodman and Burger reasoned that they might be less likely to show this defensive attribution

effect. That is, they would be less likely to defend themselves against thinking that there is a possibility that it could happen to them.

Rodman and Burger tested 56 college students in a 2×3 complex design. The first independent variable was the severity of the described accident (severe and nonsevere), and this was manipulated using the random groups design. The natural groups design was used for the second independent variable; students were selected on the basis of their scores on a paper-and-pencil test of depression to represent three groups: nondepressed, slightly depressed, and mildly depressed individuals. The dependent variable was a single item on a longer questionnaire that asked students to divide 100% among four potential sources of responsibility: each of the three drivers in the accident and "uncontrollable circumstances." The defensive-attribution effect would be reflected in a larger value assigned to uncontrollable circumstances (i.e., "chance") for the nonsevere than for the severe accident. (Remember that it is protective to avoid making attributions to uncontrollable circumstances for severe accidents.) Rodman and Burger hypothesized that the defensive-attribution effect would decrease as a person's level of depression increased.

The mean percentage values for this uncontrollable factor for each of the six conditions are presented in Table 13.5. An analysis of variance summary table for a complex design with two independent variables includes four sources of variation: the main effects of each independent variable, the interaction effect of the two independent variables, and the within-group error (see Challenge Question 3, p. 455, for an illustration). For example, in the Rodman and Burger experiment there could be a main effect of Type of Accident and a main effect of Level of Depression. The effect of primary interest in the Rodman and Burger experiment, however, was the predicted interaction effect between the two independent variables. Moreover, an ANOVA revealed that this interaction effect was statistically significant. As you can see in Table 13.5, the interaction effect arises because the differences between the percentage values for severe and nonsevere accidents changed as the degree of depression increased. Specifically, nondepressed students showed the defensive-attribution effect and mildly depressed students did not.

Once we have confirmed that there is an interaction of two independent variables, we must locate more precisely the source of that interaction effect. There are statistical tests specifically designed for tracing the source of a significant interaction effect. These tests are called simple main effects and comparisons of two means (see Keppel, 1991).

TABLE 13.5 MEAN PERCENTAGE OF RESPONSIBILITY ATTRIBUTED TO UNCONTROLLABLE CIRCUMSTANCES

	Level of depression		
Type of accident	Nondepressed	Slightly depressed	Mildly depressed
Severe	7.00 (9.2)	14.00 (16.1)	16.90 (16.0)
Nonsevere	30.50 (22.2)	16.50 (12.7)	3.75 (3.5)

Note: Standard deviations appear in parentheses. Adapted from Rodman and Burger (1985).

A *simple main effect* is the effect of one independent variable at one level of a second independent variable. In fact, one definition of an interaction effect is that the simple main effects across levels are different. We can illustrate the use of simple main effects by returning to the results of the Rodman and Burger experiment. There are five simple main effects in Table 13.5. Three of the simple main effects are represented by the effect of the type of accident (severe, nonsevere) at each of the three levels of depression considered separately. The other two simple main effects are represented by the effect of depression level (nondepressed, slightly depressed, mildly depressed) at each of the two levels of accident type. The defensive-attribution effect refers to the difference between the means for severe and nonsevere accidents, and Rodman and Burger predicted that this effect of the type of accident would decrease as the level of depression increased. Therefore, it was appropriate for these researchers to analyze the simple main effect of accident type at each level of depression to test their prediction. They found, as predicted, that the simple main effect of the type of accident was statistically significant for nondepressed students. That is, this group of nondepressed students demonstrated different attributions to uncontrollable (chance) circumstances for severe and nonsevere accidents. In contrast, the simple main effects for type of accident were not statistically significant for the slightly depressed and the mildly depressed students. Thus, for the two groups of slightly and mildly depressed students, there was no difference in their attribution to chance for severe and nonsevere accidents.

Two of the simple main effects in the Rodman and Burger (1985) study each involve three means. One can also examine how nondepressed, slightly depressed, and mildly depressed students differed in their attributions for severe accidents and how these three groups differed in their attributions for nonsevere accidents. That is, if statistical analysis reveals a significant simple main effect, then one concludes that there is a difference among the means (e.g., among the three groups for severe accidents). The next step, then, is to conduct comparisons of two means to analyze simple main effects more fully (see Keppel, 1991). That is, once a simple main effect involving more than two levels of a variable has been shown to be statistically significant, comparisons of two means can be done to determine the nature of the differences among the levels. In this procedure, means within the simple main effect are compared two at a time in order to identify the source of differences among levels. As you can see, two-mean comparisons make sense only when there is a simple main effect for an independent variable with three or more levels. With two levels, a simple main effect compares the difference between the two means and no additional comparisons are necessary.

Once an interaction effect has been thoroughly analyzed, researchers can also examine the main effects of each independent variable. In general, however, main effects are less interesting when an interaction effect is statistically significant. For example, in the Rodman and Burger (1985) experiment, we learned that the effect of the type of accident on participants' attributions to uncontrollable circumstances depended on individuals' level of depression. Based on the

analyses of main effects, we do not learn much more of interest when we add that, overall, participants attributed a higher mean percentage of responsibility to uncontrollable circumstances for nonsevere accidents compared to severe accidents. Nonetheless, there are experiments in which the interaction effect and the main effects are all of interest.

Analysis with No Interaction Effect

- If an omnibus analysis of variance indicates the interaction effect between independent variables is not statistically significant, the next step is to determine whether the main effects of the variables are statistically significant.
- The source of a statistically significant main effect can be specified more precisely by performing comparisons that compare means two at a time and by constructing confidence intervals.

When the interaction effect is not statistically significant, the next step is to examine the main effects of each independent variable. If the overall main effect for an independent variable is not statistically significant, then there is nothing more to do. However, if a main effect is statistically significant, there are several approaches a researcher may take. For example, if there are three or more levels of the independent variable, the source of a statistically significant main effect can be specified more precisely by performing comparisons of two means. (Of course, if there are but two levels, an additional comparison is not needed.) Yet another approach is to construct confidence intervals around the group means (see Chapter 12). You may see that these analyses are similar to the analyses we described for a single-factor independent groups design. The difference for the complex design is that the data for one independent variable are collapsed across the levels of other independent variables.

Effect Sizes for Two-Factor Design with Independent Groups

A common measure of effect size for a complex design using ANOVA is eta squared (η^2), or proportion of variance accounted for, which was discussed earlier in the context of single-factor designs. In calculating eta squared, it is recommended that we focus only on the effect of interest (see Rosenthal & Rosnow, 1991). Specifically, eta squared can be defined as

$$\eta^2 = \frac{SS_{\text{effect of interest}}}{SS_{\text{effect of interest}} + SS_{\text{within}}} \quad \text{(see Rosenthal \& Rosnow, 1991, p. 352)}$$

Thus, eta squared may be obtained for each of the three effects in an A × B design.

As noted above (and see Rosenthal & Rosnow, 1991), when the sums of squares for the effects are not available, eta squared can be computed using the formula

$$\eta^2 = \frac{(F)(df_{\text{effect}})}{(F)(df_{\text{effect}}) + (df_{\text{error}})}$$

Role of Confidence Intervals in the Analysis of Complex Designs

The analysis of a complex design can be aided by the construction of confidence intervals for the means of interest. For example, each mean in a 2 × 3 design can be bracketed with a confidence interval following the procedures outlined in Chapter 12 and earlier in this chapter. Recall that the formula is

$$\text{Upper limit of 95\% confidence interval: } \overline{X} + [t_{.05}][s_{\overline{X}}]$$

$$\text{Lower limit of 95\% confidence interval: } \overline{X} - [t_{.05}][s_{\overline{X}}]$$

When sample sizes are equal, the estimated standard error is defined as

$$s_{\overline{X}} = \frac{s_{\text{pooled}}}{\sqrt{n}} \quad \text{where } n = \text{sample size for each group}$$

Because the square root of the MS_{error} from the ANOVA Summary Table is equivalent to s_{pooled}, we can define the 95% confidence interval as

$$95\%\ CI = \overline{X} \pm (t_{.05})\left[\sqrt{(MS_{\text{error}}/\sqrt{n})}\right]$$

where $t_{.05}$ is defined by the degrees of freedom associated with the MS_{error}.

To illustrate the use of confidence intervals in the analysis of a complex design, we will use the results of a hypothetical 2 × 3 independent groups design. Suppose that in this experiment participants are asked to perform a motor task with their dominant hand or their nondominant hand. Furthermore, one third of the group of participants are randomly assigned to a 60-second delay between trials, another third to a 30-second delay, and the final third of the participants are assigned to a 0-second delay between trials. The dependent variable is the number of correct responses on the motor task. Variable A is Hand Dominance with 2 levels (dominant and nondominant hand). Variable B is Delay Between the Trials with 3 levels (0, 30, and 60 seconds). The means for each of the six groups in this experiment are presented below. There were 5 participants in each group ($n = 5$). It may be helpful to graph the means to reveal the nature of the interaction effect. The omnibus F-test revealed that the interaction effect was statistically significant, $F(2, 24) = 12.34$, $p < .0005$ (as were both main effects).

		Delay (B)		
		0 sec (b_1)	30 sec (b_2)	60 sec (b_3)
Hand (A)	Dom (a_1)	19.0	19.0	20.0
	NonDom (a_2)	10.6	15.8	18.2

The MS_{error} for this hypothetical experiment was 2.45 ($df = 24$). Thus, the pooled standard error of the mean is equal to

$$\left(\sqrt{MS_{\text{error}}}/\sqrt{n}\right) = \left(\sqrt{2.45}/\sqrt{5}\right) = \sqrt{.49} = .70$$

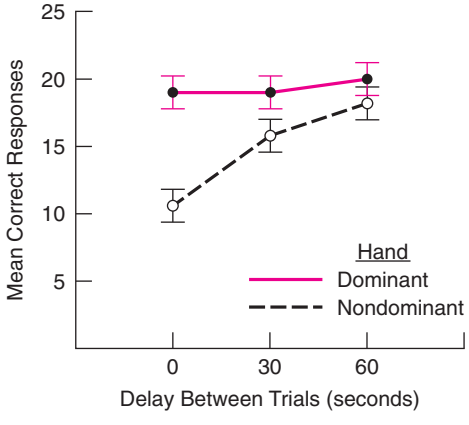
Consulting Appendix Table A.2, we find that the critical *t* value for 24 *df* is 2.06. The 95% *CI*s for the six groups in this experiment are

$$a_1b_1 = \overline{X} \pm (2.06)(.70) = 19.0 \pm 1.44 = 17.56 - 20.44$$
$$a_1b_2 = \overline{X} \pm (2.06)(.70) = 19.0 \pm 1.44 = 17.56 - 20.44$$
$$a_1b_3 = \overline{X} \pm (2.06)(.70) = 20.0 \pm 1.44 = 18.56 - 21.44$$
$$a_2b_1 = \overline{X} \pm (2.06)(.70) = 10.6 \pm 1.44 = 9.16 - 12.04$$
$$a_2b_2 = \overline{X} \pm (2.06)(.70) = 15.8 \pm 1.44 = 14.36 - 17.24$$
$$a_2b_3 = \overline{X} \pm (2.06)(.70) = 18.2 \pm 1.44 = 16.76 - 19.64$$

Figure 13.3 shows the confidence intervals around the six means in the hypothetical experiment. An examination of the *CI*s tells us about the precision of our estimates. We want to examine the interval width and the probable pattern of *population* means by looking to see if the intervals around the sample means overlap and, if so, to what degree they overlap. Recall that a rule of thumb for interpreting confidence intervals suggests that if the intervals around means do not overlap, then the two means would likely be statistically significant if tested using NHST (see Box 12.5 in Chapter 12).

Based on these 95% confidence intervals, the interaction effect indicates that the effect of the hand-dominance variable varies depending on the level of the delay variable. In the 0-second and 30-second delay conditions, the number of correct responses is greater for the dominant hand compared to the nondominant hand. This difference in motor-task performance when comparing dominant and nondominant hands is greatest for the 0-second delay between trials, followed by the 30-second delay condition. There may be no difference in motor-task performance between groups using the dominant and the nondominant hand when there is a 60-second delay between trials (i.e., the confidence intervals overlap); additional data are needed before a firm conclusion may be made.

FIGURE 13.3 Mean number of correct responses as a function of the delay between trials (in seconds) and hand used to perform motor task (dominant, nondominant). The 95% confidence interval is shown around each mean.



	0	30	60
D	19	19	20
ND	10.6	15.8	18.2

CIs (bars) are ±1.44 units

The second way to explain the interaction effect is to describe the effect of the delay variable for each level of hand dominance. For the dominant hand condition, the 95% confidence intervals overlap across the three levels of time delay, and each of the three sample means is included within the three confidence intervals representing the delay levels. Thus, we can reasonably conclude that the delay variable does not affect motor-task performance for participants who use their dominant hand. In contrast, the number of correct responses within the nondominant hand condition varies as a function of the delay variable. The 95% confidence intervals indicate fewer correct responses in the 0-second condition compared to the 30-second and the 60-second delay conditions, but that performance in these latter two conditions may not differ (the results are inconclusive).

Two-Factor Analysis of Variance for a Mixed Design

The two-factor analysis of variance for a mixed design is appropriate when one independent variable represents either the random groups or natural groups design and the second independent variable represents the repeated measures design. The first independent variable is called the between-subjects factor (here symbolized as A). The second independent variable is called the within-subjects factor (symbolized as B). The two-factor analysis for a mixed design is somewhat of a hybrid of the single-factor analysis for independent groups and the single-factor analysis for the repeated measures designs. This particular complex design was not discussed in Chapter 9 and, therefore, we will present it in some detail here.

The data presented in the following table represent the mean frequency judgments participants gave to four brief segments of popular songs. The

		Presentation frequency (B)		
Group (A)	Subject	1 (b_1)	2 (b_2)	3 (b_3)
Incidental	1	1.2	2.2	3.0
	2	.8	2.0	3.2
	3	1.2	1.8	2.8
group (a_1)	4	1.5	2.0	2.5
	5	1.2	2.5	3.2
	M	1.18	2.1	2.94
	SD	.25	.26	.30
	Range	.8–1.5	1.8–2.5	2.5–3.2
Intentional	6	1.2	1.5	2.0
	7	1.0	2.0	2.5
	8	1.0	1.2	3.2
group (a_2)	9	1.2	2.2	2.8
	10	1.0	2.2	3.5
	M	1.08	1.82	2.80
	SD	.11	.45	.59
	Range	1.0–1.2	1.2–2.2	2.0–3.5

participants listened to a tape including many songs; half the individuals did not expect the frequency judgment test (incidental group), and half did expect the test (intentional group). In addition, all participants judged the frequency of songs that had been presented either one, two, or three times. Thus, the experiment was a 2 × 3 design in which instructions were manipulated in a random groups design with five people assigned to each of two groups and in which the presented frequency variable was manipulated in a complete repeated measures design. The following data matrix shows the mean frequency judgments at three levels of presentation frequency for each participant in each group, as well as summary statistics in the form of means, standard deviations, and ranges, for each condition.

As you should now be aware, it is important to appreciate the trends in the data before looking at the ANOVA Summary Table. For instance, note that the mean frequency judgments increase in each group as a function of the presented frequency. Interestingly, there is not that great a difference between group means (i.e., incidental vs. intentional) within each frequency presentation condition. An outline of typical computer output for a two-factor analysis of variance for a mixed design is presented below; however, be aware that some computer programs separate the output of a mixed design, showing first the output for the between-groups analysis and then the output for the within-subjects analysis (which includes the interaction effect). You may find that you have to scroll the computer screen to get all of the information.

		Between subjects			
Source	***SS***	***df***	***MS***	***F***	***p***
Group	0.225	1	0.225	1.718	0.226
Error	1.049	8	0.131		
		Within subjects			
Present	15.149	2	7.574	58.640	0.000
Present × Group	0.045	2	0.022	0.173	0.843
Error	2.067	16	0.129		

The summary table is divided into two parts. The "Between subjects" section includes the *F*-ratio for the main effect of groups. The form of this part of the table is like that of a single-factor analysis for the independent groups design. The error listed in this section is the within-groups variation. The *F*-test for the effect of group was not statistically significant because the obtained probability of .226 was greater than the conventional level of statistical significance of .05. The second part of the summary table is headed "Within subjects." It includes the main effect of the within-subjects variable of presentation frequency and the interaction of presentation frequency and group. In general, any effect including a within-subjects variable (main effect or interaction effect) must be tested with the residual error term used in the within-subjects design. The *F*-test for the interaction effect is less than 1, so was not statistically significant. The main effect of presentation frequency, however, did result in a statistically significant *F*. (As was true in the

analysis of the single-factor within-subjects design, your computer output may include additional information beyond what we have presented here.)

Interpreting the results of a two-factor analysis for a mixed design follows the logic for any complex design. That is, the interaction effect is examined first, and if the interaction effect is statistically significant, then simple main effects and comparisons of two means are used to identify the source of the interaction effect. The main effects are then examined, and two-mean comparisons can be used to analyze further statistically significant main effects of independent variables with more than two levels.

Care must be taken when analyzing a mixed design to use the appropriate error term for analyses beyond those listed in the summary table (i.e., simple main effects, comparisons of two means). For example, if a significant interaction effect is obtained, it is recommended that simple main effects be analyzed by treating each simple effect as a single-factor ANOVA at that level of the second independent variable. If, for instance, we had obtained a significant interaction effect between group and presentation frequency in our sample experiment, a simple main effect for the intentional group would involve carrying out a repeated measures ANOVA for only that group (see Keppel, 1991, for more information on these comparisons).

Effect size estimates in a mixed design also frequently make use of eta squared, that is, an estimate of proportion of variance accounted for by the independent variable. As in the independent groups design (see above), eta squared is defined as the *SS* effect divided by the *SS* effect plus the *SS* error for that effect.

Reporting Results of a Complex Design

Reporting results of a complex design follows the general form of a report for a single-factor ANOVA but gives special attention to the nature of an interaction effect when it is present. The following are important elements of a report of the results of a complex design:

—description of variables and definition of levels (conditions) of each;
—summary statistics for cells of the design matrix in text, table, or figure, including when appropriate, confidence intervals for group means;
—report of *F*-tests for main effects and interaction effect with exact probabilities;
—effect size measure for each effect;
—statement of power for nonsignificant effects;
—simple main effects analysis when interaction effect is statistically significant;
—verbal description of statistically significant interaction effect (when present), referring reader to differences between cell means across levels of the independent variables;
—verbal description of statistically significant main effect (when present), referring reader to differences among cell means collapsed across levels of the independent variables;
—comparisons of two means, when appropriate, to clarify sources of systematic variation among means contributing to main effect;
—conclusion that you wish reader to make from the results of this analysis.

An example of a Results section can be found in the Sample Research Report in Chapter 14.

SUMMARY

Statistical tests based on null hypothesis significance testing (NHST) are commonly used to perform confirmatory data analysis in psychology. NHST is used to determine whether differences produced by independent variables in an experiment are greater than what would be expected solely on the basis of error variation (chance). The null hypothesis is that the independent variable did not have an effect. A statistically significant outcome is one that has a small probability of occurring if the null hypothesis were true. Two types of errors may arise when doing NHST. A Type I error occurs when a researcher rejects the null hypothesis when it is true. The probability of a Type I error is equivalent to alpha or the level of significance, usually .05. A Type II error occurs when a false null hypothesis is not rejected. Type II errors can occur when a study does not have enough power to correctly reject a null hypothesis. The primary way researchers increase power is by increasing sample size. By using power tables researchers may estimate, before a study is conducted, the power needed to reject a false null hypothesis and, after a study is completed, the likelihood of detecting the effect that was found. The exact probability associated with the result of a statistical test should be reported.

The appropriate statistical test for comparing two means is the *t*-test. When the difference between two means is tested, an effect size measure, such as Cohen's *d*, should also be reported. The APA *Publication Manual* strongly recommends that confidence intervals be reported as well as the results of NHST. When reporting the results of NHST, it is important to keep in mind that statistical significance (or nonoverlapping confidence intervals) is not the same as scientific or practical significance. Moreover, neither NHST, confidence intervals, nor effect sizes, tell us about the soundness of a study's methodology. That is, none of these measures alone may be used to state that the alternative hypothesis (that the independent variable did have an effect) is correct. Only after we have examined carefully the methodology used to obtain the data for an analysis will we want to venture a claim about what influenced behavior.

Analysis of variance (ANOVA) is the appropriate statistical test when comparing three or more means. The logic of ANOVA is based on identifying both error variation and sources of systematic variation in the data. An *F*-test is constructed that represents error variation and systematic variation (if any) divided by error variation alone. Results of the overall analysis, called an omnibus *F*-test, are reported in an ANOVA Summary Table. A large *F*-ratio provides evidence that the independent variable had an effect. Effect size measures for a single-factor independent groups design include Cohen's *f* and eta squared (η^2). Comparisons of two means may be conducted following results of an omnibus *F*-test in order to more clearly specify the sources of systematic variation contributing to a significant omnibus *F*-test. Confidence intervals, too, may be meaningfully used to complement an ANOVA conducted with

data from a multiple-group study and should be reported when the results of NHST are summarized.

A two-factor ANOVA is appropriate when a researcher examines simultaneously the effect on behavior of two or more independent variables in a complex design. When one independent variable represents an independent groups variable (random or natural groups) and another is a repeated measures within-subjects variable, we speak of a mixed design. An omnibus *F*-test is carried out to assess both main effects and the interaction effect of variables. When a statistically significant interaction effect is found, the source of the interaction effect may be pursued by conducting simple main effects. A simple main effect is the effect of an independent variable at only one level of a second independent variable. Confidence intervals, too, may be used to help understand the effect of an independent variable in a complex design. A commonly used measure of effect size in a complex design is eta squared.

Key Concepts

null hypothesis (H_0) 415
level of significance 415
Type I error 417
Type II error 417
power 418
***t*-test for independent groups 421**
repeated measures (within-subjects) *t*-test 422
ANOVA 427
single-factor independent groups design 428
***F*-test 428**
omnibus *F*-test 429
eta squared (η^2) 433
Cohen's *f* 434
comparison of two means 435

Review Questions

1. What does it mean to say that the results of a statistical test are "statistically significant"?
2. Differentiate between Type I and Type II errors as they occur when carrying out NHST.
3. What three factors determine the power of a statistical test? Which factor is the primary one that researchers can use to control power?
4. Why is a repeated measures design likely to be more sensitive than a random groups design?
5. Describe one advantage and one limitation of using measures of effect size.
6. Why may a statistically significant result be neither scientifically nor practically significant?
7. Outline briefly the logic of the *F*-test.
8. Distinguish between the information you gain from an omnibus *F*-test and from comparisons of two means.
9. What is the primary way that a repeated measures ANOVA differs from that of an ANOVA for independent groups?
10. How does a simple main effect differ from an overall main effect?

CHALLENGE QUESTIONS

1 A researcher conducts an experiment comparing two methods of teaching young children to read. An older method is compared with a newer one, and the mean performance of the new method was found to be greater than that of the older method. The results are reported as $t(120) = 2.10$, $p = .04 (d = .34)$.

- **A** Is the result statistically significant?
- **B** How many participants were there in this study?
- **C** Based on the effect size measure, *d*, what may we say about the size of the effect found in this study?
- **D** The researcher states that on the basis of this result the newer method is clearly of practical significance when teaching children to read and should be implemented right away. How would you respond to this statement?
- **E** What would the construction of confidence intervals add to our understanding of these results?

2 A social psychologist compares three kinds of propaganda messages on college students' attitudes toward the war on terrorism. Ninety ($N = 90$) students are randomly assigned in equal numbers to the three different communication conditions. A paper-and-pencil attitude measure is used to assess students' attitudes toward the war after they are exposed to the propaganda statements. An ANOVA is carried out to determine the effect of the three messages on student attitudes. Here is the ANOVA Summary Table:

Source	Sum of squares	*df*	Mean square	*F*	*p*
Communication	180.10	2	90.05	17.87	0.000
Error	438.50	87	5.04		

- **A** Is the result statistically significant? Why or why not?
- **B** What effect size measure can be easily calculated from these results? What is the value of that measure?
- **C** How could doing comparisons of two means contribute to the interpretation of these results?
- **D** Although the group means are not provided, it is possible from these data to calculate the width of the confidence interval for the means based on the pooled variance estimate. What is the width of the confidence interval for the means in this study?

3 A developmental psychologist gives 4th-, 6th-, and 8th-grade children two types of critical thinking tests. There are 28 children tested at each grade level; 14 received one form (A or B) of the test. The dependent measure is the percentage correct on the tests. The mean percentage correct for the children at each grade level and for the two tests is as follows:

Test	4th	6th	8th
Form A	38.14	63.64	80.21
Form B	52.29	68.64	80.93

Here is the ANOVA Summary Table for this experiment:

Source	Sum of squares	*df*	Mean square	*F*	*p*
Grade	17698.95	2	8849.48	96.72	.000
Test	920.05	1	920.05	10.06	.002
Grade × Test	658.67	2	329.33	3.60	.032
Error	7136.29	78	91.49		

- **A** Draw a graph showing the mean results for this experiment. Based on your examination of the graph, would you suspect a statistically significant interaction effect between the variables? Explain why or why not.
- **B** Which effects were statistically significant? Describe verbally each of the statistically significant effects.
- **C** What are the eta-squared values for the main effects of grade and test?
- **D** What further analyses could you do to determine the source of the interaction effect?
- **E** What is the simple main effect of Test for each level of Grade?
- **F** Calculate confidence intervals for the six means in the experiment, and draw them around the means in your graph of these results.

Answer to Stretching Exercise

Statements 1 and 5 are True; 2, 3, and 4 are False.

Answer to Challenge Question 1

A Yes. The obtained probability of this result assuming the null hypothesis is true is less than .05, the conventional level of significance.

B The degrees of freedom (*df*) are reported to be 120. For an independent groups *t*-test, $df = n_1 + n_2 - 2$. Thus, there must have been 122 participants.

C Cohen's guidelines suggest that an effect size of .20 is a small effect, .50 a medium or average effect, and .80 a large effect. An effect size of .34 is a small effect.

D The results of NHST do not speak directly to practical significance. If the newer method is much more expensive, too time-consuming to implement, or requires resources (e.g., new reading materials) that are not immediately available, then the practical significance of this finding (at least in the short run) is likely to be small. This may be especially the case because the effect size is rather small. Also, the fact that $p = .04$ suggests that the probability of replicating this statistically significant finding at the .05 level is not that high. Finally, we would want to examine carefully the methodology of the study to determine that the study was sound, free of confounds and experimenter errors.

E Constructing a confidence interval for the difference between the two population means would provide evidence of the size of the difference between these methods and indicate (based on examining the width of the interval) the precision of the estimation of the difference between two population means.