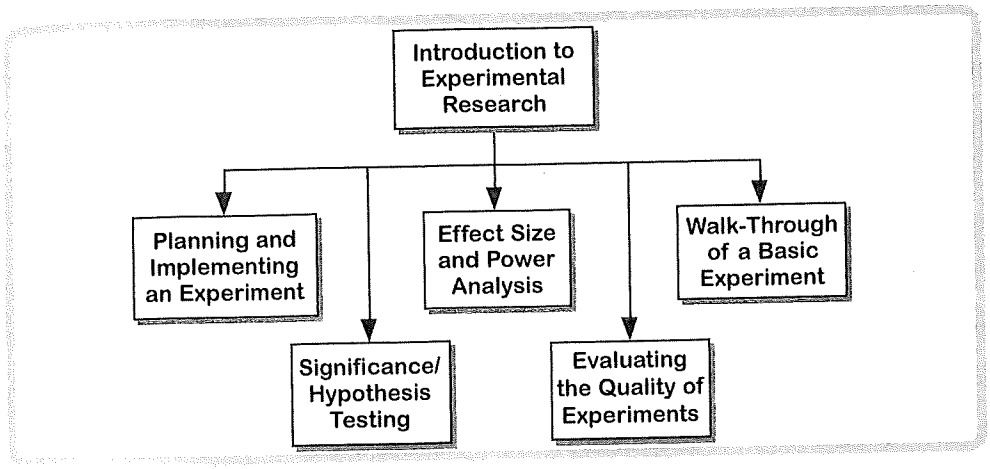
CHAPTER EIGHT

Introduction to Experimental Research



Experimental research depends on a researcher to control all aspects of a study from start to finish. This chapter describes how this is done in experimental research in general, and it explains the most basic experimental design.

CHAPTER OBJECTIVES

- ✓ Understand that the purpose of an experiment is to investigate a causal relationship.
- ✓ Understand the role of the researcher: to control all aspects of the experiment from start to finish.
- ✓ Understand the research review and how it identifies a theory that supports a research hypothesis.
- ✓ Understand independent and dependent variables.
- ✓ Understand the unit of analysis, sampling, data collection, and data analysis in experiments.
- ✓ Understand the difference between a true experiment and a quasi-experiment.
- ✓ Understand treatment and control groups and how they are used in between-group experiments.
- ✓ Understand inferential statistics in significance and hypothesis testing.
- ✓ Understand one- and two-tailed tests.
- ✓ Understand the *t*-test and how it is used in between-group experiments.
- ✓ Understand Type I and Type II errors in making decisions about hypotheses. Understand effect size estimates and power analysis.
- ✓ Understand the concept of validity.
- ✓ Understand how to evaluate an experiment.

PLANNING AND IMPLEMENTING AN EXPERIMENT MAP 8.1

In planning an experiment, researchers have to (1) review relevant research to support the experiment, and (2) make decisions about elements of design.

Researchers review relevant and recent research in order to make a case for the study and to develop a rationale for hypothesizing that a particular intervention or treatment will effect a desired change in subjects. A well-grounded intervention has a strong explanatory theory that can lead to the identification of variables and the statement of a research hypothesis.

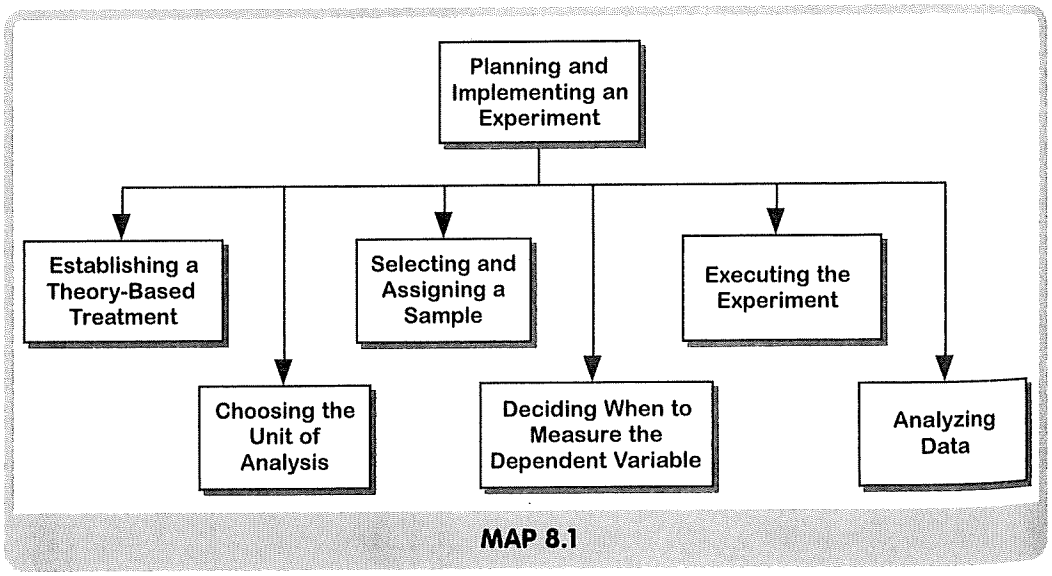
An *independent variable (IV)* represents the causal agent of the theory. Depending on their level of complexity, experimental studies may have one or multiple IVs (see Figure 8.1).

A *dependent variable (DV)* represents the outcome, or the *effect* that the researcher predicts will occur as a result of the intervention (see Figure 8.1).

The *research hypothesis* predicts the effect of the IV(s) on the DV(s). It can be stated as either a directional or a nondirectional hypothesis.

A *directional hypothesis* predicts that the treatment will result in a change, and that the change will be a positive result of the experiment.

A *nondirectional hypothesis* predicts that a treatment will result in a change in outcomes, but does not predict the direction of the change (i.e., whether it will be positive or negative).



MAP 8.1

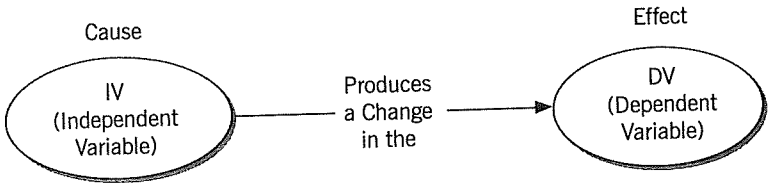


FIGURE 8.1. The relationship of the independent variable (IV) and the dependent variable (DV).

Choosing the Unit of Analysis

Between-group designs compare outcomes between groups after an intervention has occurred.

The *treatment group* receives the theory-based intervention.

The *control group* continues with "business as usual" and serves as a reference point for comparison with the treatment group.

Repeated-measures designs compare outcomes for individual subjects within a group.

Repeated-measures designs, and more complex types of between-group designs, are discussed in Chapter Nine. This chapter focuses on the simplest type of between-group design, with two groups and a single intervention.

Selecting and Assigning a Sample

Ideally, an experimental researcher will randomly select a sample of 60 or more subjects from a population. While randomized selection makes it more likely that findings can be generalized from the sample to the population, it is often not feasible for researchers to do this. In these cases, the researchers will use a convenience sample.

In a between-group design, researchers also have to decide how to assign the sample to control and treatment groups. The method of group assignment is what distinguishes a *true experimental design* from a *quasi-experimental design*.

A *true experimental design* uses a table of random numbers to assign subjects to groups.

A *quasi-experimental design* assigns subjects to groups nonrandomly, and usually assigns different *intact groups* (e.g., classrooms and schools) to treatment and control groups.

Deciding When and How Often to Measure the DV

Data may be collected on the DV before an intervention is introduced *and* after the experiment has concluded, or they may be collected only after an experiment has concluded.

A *pre- and posttest design* measures the DV(s) before the IV is introduced and again at the conclusion of the experiment. The most basic pre- and posttest design has one treatment group and one control group.

A *posttest-only design* measures the DV(s) only at the conclusion of the experiment.

Executing the Experiment

The researcher manipulates the IV so that the theory-based treatment is administered and is compared with a control condition. The researcher tries, as much as possible, to control all conditions of the experiment. However, this is more difficult to do with human subjects in real-life situations than with nonhuman subjects in a laboratory. The experiment concludes with a final measurement of the DV(s).

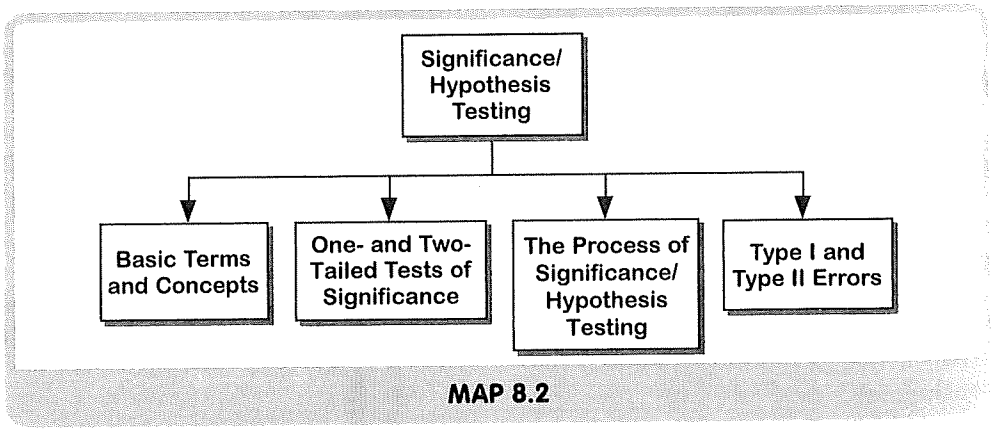
Analyzing Data

Once researchers have collected data from the posttest, they either calculate the means and standard deviations or the variances on the posttest. When they find changes in the DV, they use a process of statistical reasoning to establish statistical significance and test whether a hypothesis is true.

SIGNIFICANCE/HYPOTHESIS TESTING MAP 8.2

Basic Terms and Concepts

The purpose of significance/hypothesis testing is to determine the probability (expressed as a statistic, p) that findings are due to an error in sampling. A low p-value indicates that the findings are *not* due to sampling error and are most likely due to the treatment. This allows the researcher to confirm the hypothesis.



MAP 8.2

Statistical significance means that there is a low probability of the results being due to sampling error. It does not mean that the findings are important. The lower the *p*-value, the better.

Alpha (α) is the term used to indicate the level of *p* the researcher will accept for statistically significant findings. The researcher establishes the alpha before the experiment begins.

$p \leq .05$ is the conventional alpha for experiments in education and other social sciences. It means a 5% or lower probability that the results are due to sampling error and a much greater likelihood that they are due to the treatment.

Significance/hypothesis testing in experiments depends on inferential statistics, sometimes referred to as *tests of significance.*

The *t-test* is the simplest way to establish the statistical significance of results between two groups. The *t*-test yields a statistic, *t*, which is then aligned with a *p*-value.

Analysis of variance (ANOVA) is a more robust statistic because it can be used with two or more groups and with additional IVs. *Multiple analysis of variance* (MANOVA) extends ANOVA to combine two or more DVs in the analysis. ANOVA/MANOVA yield a statistic, *F*, which is aligned with a *p*-value.

Analysis of covariance (ANCOVA) adjusts values on the DV when there is reason to think that the groups being compared are not equivalent. Nonequivalence can be detected on a pretest; it may be assumed in quasi-experiments when nonrandom assignment may result in uneven groups; or it may be used to control for another variable. *Multiple analysis of covariance* (MANCOVA) extends ANCOVA to combine two or more DVs in the analysis.

One- and Two-Tailed Tests of Significance

All inferential tests are based on probabilities of there being errors in sampling; these probabilities are derived from the bell curve and from a table of probabilities with the sample size and level of probability to guide the decision. Researchers can choose between using a *one-tailed* or *two-tailed* inferential test.

A *one-tailed test* (Figure 8.2) uses one end, or tail, of the curve to generate $p \leq .05$.

A *two-tailed test* (Figure 8.3) uses two ends, or tails, with $p \leq .025$ on each end of the curve, to generate $p \leq .05$.

A two-tailed test is the more objective type of test. It represents a higher standard and degree of difficulty, because the probability must be distributed to both ends of the curve. Researchers often use a two-tailed test when there is a nondirectional hypothesis. A one-tailed test is more lenient, because all of the probability for the hypothesis test is on one side of the distribution curve.

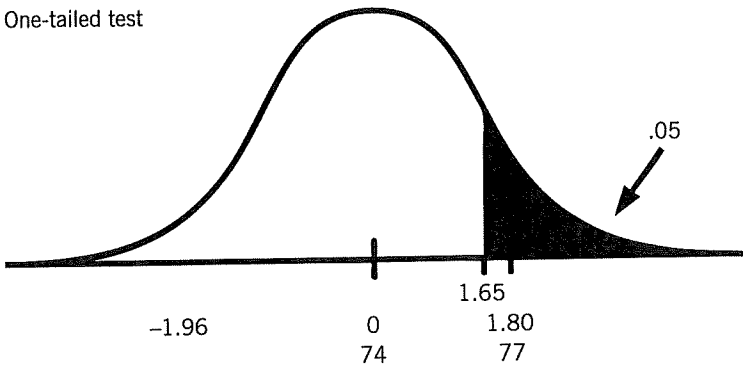


FIGURE 8.2. One-tailed test.

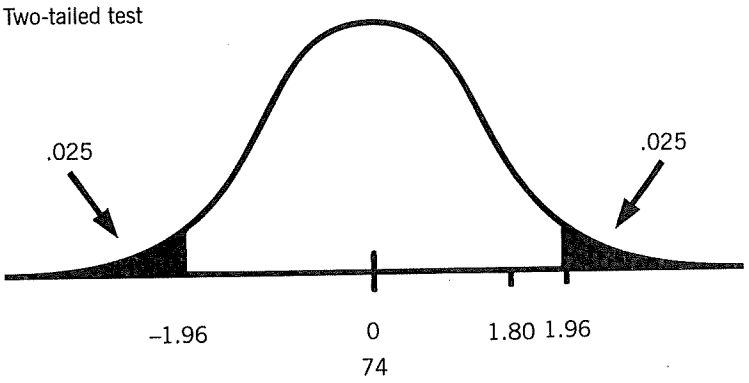


FIGURE 8.3. Two-tailed test.

The Process of Significance/Hypothesis Testing

Significance/hypothesis testing is a multistep statistical reasoning process that involves the *null hypothesis.*

The *null hypothesis* restates the hypothesis in negative terms *and* predicts that there will be no significant difference on the DV between the treatment and control groups.

To confirm that a hypothesis is true, a researcher has to demonstrate there is a low probability that the null hypothesis is true and that results are due to sampling error. In effect, the null hypothesis is set up as a "straw man"; it is easier to prove something false than it is to prove something true. This is where inferential statistics come in. The *p*-value represents the probability that results are due to error and that the null is true.

Before the experiment begins, the researcher establishes an *alpha* (α), which identifies the level of error the researcher will accept (see p. 145). For instance, if the researcher establishes an alpha of $p \leq .05$ and significance testing yields $p \leq .05$, the researcher can conclude that there is a 5% or lower probability that the null hypothesis is true. In other words, when $p \leq .05$, the researcher can reject the null hypothesis and confirm the research hypothesis.

Type I and Type II Errors

Despite all of this attention to statistical reasoning, there is still no guarantee that the decision about the hypothesis is correct; this is an admission of research fallibility. The possibility of making a wrong decision is known as a *Type I* or *Type II* error (see Figure 8.4).

Type I error, also known as the *alpha-level error*, occurs when the researcher rejects the null hypothesis even though this hypothesis is true.

Type II error, also known as the *beta-level error*, occurs when the researcher accepts the null hypothesis even though this hypothesis is false.

A researcher can control for Type I error by simply adjusting the alpha level for hypothesis testing to $p \leq .025$, thus making it more difficult to reject the null hypothesis. Correcting for a Type II error is more complicated and requires strengthening the theory, increasing the sample size, or improving the measurement.

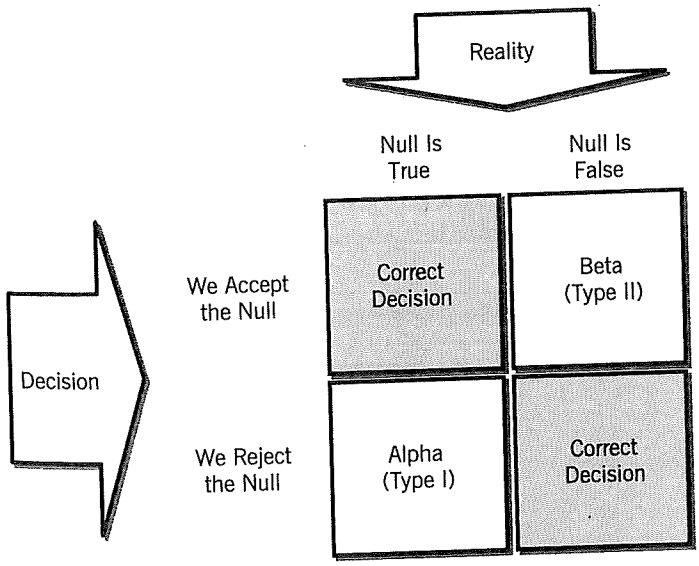


FIGURE 8.4. Decision making about the null hypothesis (with Type I and Type II errors).

EFFECT SIZE AND POWER ANALYSIS

MAP 8.3

While inferential statistical significance testing can indicate the probability that differences are due to treatments and not to chance or sampling error, it cannot indicate the magnitude of differences or the strength of an intervention. This is left to another statistic, known as the *effect size*.

Effect Size

Effect size (often abbreviated as ES, especially in equations) is a standardized number that describes the overall strength of an experimental intervention in terms of changes in the standard deviation that result from the IV.

For example, ES = 0.5 means there has been a one-half standard deviation increase in values due to the treatment.

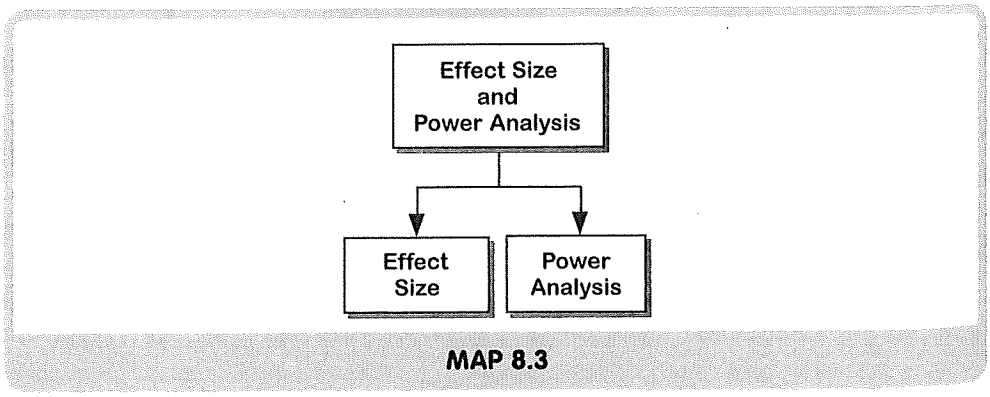
There are several formulas for determining effect size. For instance, the simplest delta (Δ) formula, developed by Glass (1976), is shown below. More advanced and sophisticated formulae are also used.

$$\Delta \text{ (delta)} = \frac{\text{Mean (experimental group)} - \text{mean (control group)}}{SD \text{ of the control group}}$$

There are also *Cohen's d* and *Hedges's g*. The following guidelines for judging ES are recommended:

<0.1 = trivial effect
0.1–0.3 = small effect
0.3–0.5 = moderate effect
>0.5 = large effect

In most cases, ES = 0.5 is desired.



MAP 8.3

An analogy for effect size is the *signal-to-noise ratio*, which compares the size of the desired signal to extraneous or competing noise (Bickman & Rog, 2009). In addition to reporting on the strength of an intervention in an experiment, effect size estimates can be used to calculate the overall impact of an intervention across multiple studies, called *meta-analyses*. The use of effect size in meta-analysis is discussed in detail in Chapter Sixteen.

Power Analysis

Just as a researcher sets an acceptable level for p (i.e., alpha) before the experiment begins, the researcher may also set an estimated effect size. Taken together, the estimated effect size and the alpha can be used to address the age-old issue of sample size and help avoid a Type II error. This is done through the process of *power analysis*, which the researcher conducts before deciding on the final sample size for the study. An adequate sample size will yield a power = .80. To conduct a power analysis, the researcher uses a power software package and enters the following information:

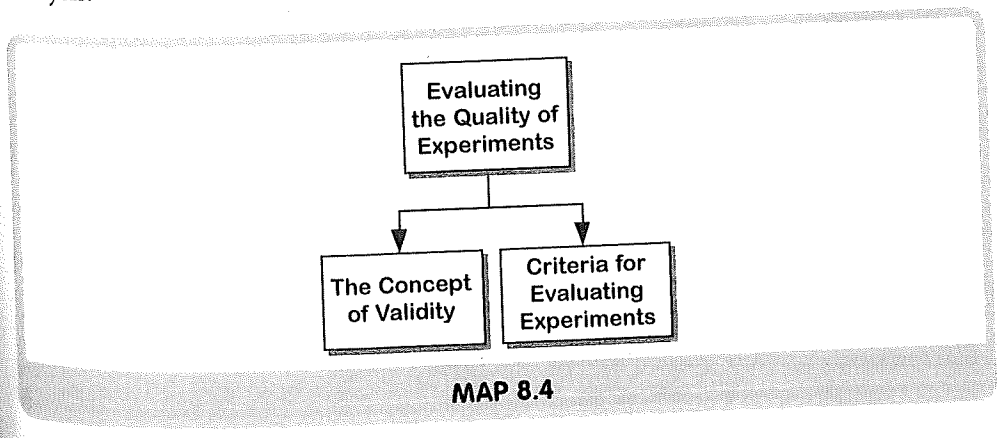
- The inferential test that will be used to analyze data.
- The alpha for significance that will be used.
- The expected effect size (again, ES = 0.5 is recommended).
- The planned sample size for the study.

The software will then generate a power value between 0 and 1. If power is less than .80, the researcher will have to increase the sample size in order to avoid a Type II error. If the sample cannot be increased, the researcher may continue with the study and acknowledge that the study has less power.

EVALUATING THE QUALITY OF EXPERIMENTS

MAP 8.4

An evaluation of the quality of an experimental study focuses on its validity and the importance of theory and treatment, sample and sampling, and data collection and analysis.



MAP 8.4

The Concept of Validity

MAP 8.5

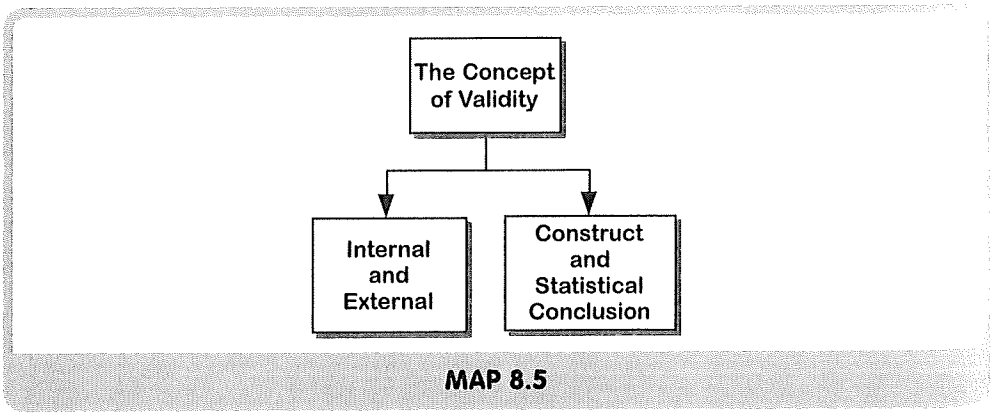
The concept of *validity* was introduced in Chapter Seven in relation to measures and the degree to which they measure what they purport to measure. In evaluating an experimental study, validity concerns the degree to which the experiment is well planned and executed, establishes a causal relationship between and among variables, and allows the researcher to generalize the results to other populations and settings. Campbell and Stanley (1963) developed the first model of validity; they identified two types of validity (*internal* and *external*) and the threats that could undermine them.

Internal validity is concerned with the strength of causality: whether the changes in the DV are due to the treatment or to alternative explanations or threats that might account for the differences.

Threats to internal validity include (1) maturation (normal developmental processes); (2) history (an unexpected event that affects the relationship between the IV and DV); (3) statistical regression (a sample that includes extremes, such as all "gifted" students); (4) selection (nonrandomized assignment of subjects, mortality, attrition of subjects); (5) testing (unintended sensitizing of subjects during pretesting); (6) design contamination (subjects' comparing notes or having motivations for the experiment to succeed or fail); and (7) compensatory rivalry (for example, teachers in a control modifying their behavior in order to compete with the treatment group).

External validity is concerned with the strength of generalizability: the degree to which findings can be applied to other individuals, settings, and times (*population validity*) or to other settings (*ecological validity*).

Threats to population validity include (1) nonrandom samples and nonrandom group assignments; (2) the Hawthorn effect (the reaction of subjects to being studied); (3) the experimental conditions effect (the way the experiment is arranged); and (4) the experimenter effect (the characteristics of the person[s] conducting the study). Threats to ecological validity include the second, third, and fourth threats to population validity.



MAP 8.5

The concept of validity was elaborated 16 years later by Cook and Campbell (1979), who added *construct validity* and *statistical conclusion validity* to the mix.

Construct validity (also known as *theoretical validity*) concerns the strength of the theory—its ability to support the choice of treatment, the operationalization of variables, and the development of a hypothesis suitable for hypothesis testing.

Statistical conclusion validity concerns the strength and precision of statistical reasoning to make inferences from the data.

Threats to statistical conclusion validity include (1) "fishing" (reanalyzing data in order to find possible significant results); (2) invalid or unreliable measures; and (3) inadequate sample size or sampling error.

Criteria for Evaluating Experiments

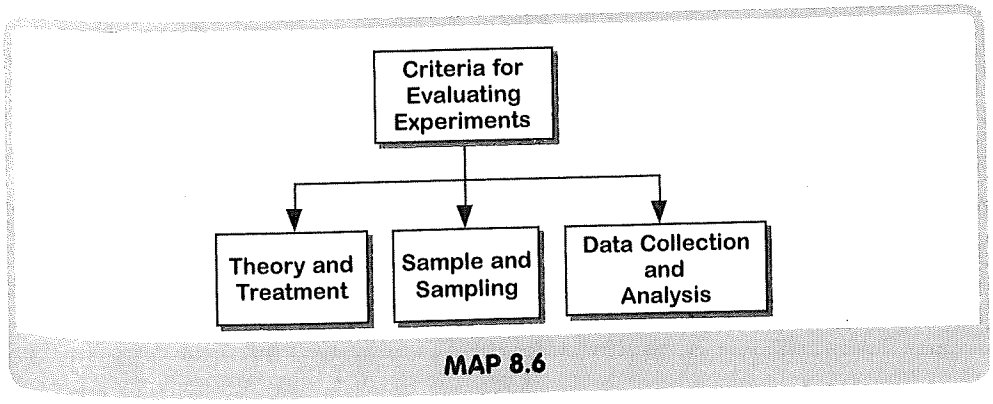
MAP 8.6

We recommend that readers evaluate experimental studies in terms of their theory and treatment, their samples and sampling, and their collection and analysis of data. We believe that these three criteria are reasonable ways to capture the complexity of validity and threats to validity.

Theory and Treatment

The primary considerations for theory and treatment are the quality of the literature review, how well the authors derived operational definitions for the independent and dependent variables, how well the treatment matches the theory, and how well the treatment was implemented.

1. How well does the research review support a theory?
2. How well is the theory operationalized? How adequate is the operationalization of the IV and DV?
3. How well does the treatment match the theory?
4. How well is the treatment implemented?



MAP 8.6

Sample and Sampling

The key elements in sampling are the procedures for selecting the sample and assigning subjects to the treatment condition, as well as the size of the sample and the characteristics of the subjects.

1. Is randomization used in selection and/or assignment to the treatment condition?
2. Is there adequate sample size? (Was there a power analysis? Does the sample include a minimum of 30 subjects?)

Collection and Analysis of Data

The key elements in evaluating collection and analysis of data are the validity and reliability of the measures, as well as the appropriateness of the statistical analysis and the process of statistical reasoning. Figure 8.5 shows a rubric for evaluating the quality of experiments.

Criterion	Strong	Moderate	Weak
Theory and treatment	Clear, current, and sufficient literature review, including 10 or more references; clearly defined IV and DV.	Literature review may not make a clear connection to the research question or has fewer than 10 references; IV and DV definitions missing or unclear.	Unclear literature review; fewer than 5 references; no reference to or definitions of IV and DV.
Sample and sampling	Clear description of the sample characteristics and the population; clear description of sample; evidence of sufficient size (60 or more, or use of power analysis); random assignment of subjects to control conditions. The sample allows for the generalization of findings to other subjects and settings.	Description of sample is present, but lacking in details about characteristics and the population; insufficient sample size; nonrandom assignment to control condition; unclear description of sample size. The sample does not allow for generalization of findings.	No description of the sample; total lack of details about the population; nonrandom assignment to control condition; insufficient sample size. The sample does not allow for generalization of findings.
Data collection and analysis	Clear description of measures selected to collect data on DV and evidence of validity and reliability (r); clear description of statistical significance (p) and of the inferential tests used; clear reporting on hypothesis testing; clear indication of whether ES was calculated; and clear reporting of results. Researcher stays close to data and does not overconclude.	Mention of measurement, but lacking in details about validity and reliability (r); mention of statistical significance (p), but without including inferential tests used; no reporting on hypothesis testing or indication of ES. Researcher avoids overconcluding.	Minimal or no mention of measurement; no mention of validity or reliability; no mention of significance, hypothesis testing, or use of ES. Researcher may overconclude.

FIGURE 8.5. Rubric for evaluating quality of experiments.

1. How good are the validity and reliability of the measures?
2. How well are the measured data reported and analyzed?
3. Are appropriate statistics applied? Statistical significance? Effect size?
4. Are hypotheses adequately tested and evaluated? Is the alpha set too high?
5. Are there alternative explanations due to extraneous and confounding factors, or threats that affect the conclusion?

WALK-THROUGH OF A BASIC EXPERIMENT

The most basic experimental designs are between-group designs with one IV, one control group, and one treatment group, as illustrated in Figures 8.6 and 8.7. This section describes the step-by-step process in a basic posttest-only experiment that a researcher used to investigate the effects of verbal praise on graduate students.

Hancock, D. R. (2002). Influencing graduate students' classroom achievement, homework habits and motivation to learn with verbal praise. *Educational Research, 44*(1), 83–95.

Research Review

An extensive research review of more than 35 studies established the efficacy of well-constructed verbal praise in improving student achievement, home preparedness, and motivation. This led to the development of three research hypotheses:

1. Postsecondary students exposed to well-administered verbal praise by a professor would demonstrate higher achievement levels on a professor-made examination than would students who received no well-administered verbal praise.
2. Postsecondary students exposed to well-administered verbal praise by a professor would spend significantly more time preparing at home (i.e., doing homework) for each lesson than would students who received no well-administered verbal praise.
3. Postsecondary students exposed to well-administered verbal praise by a professor would demonstrate higher motivation levels to learn in the classroom than would students who received no well-administered verbal praise.

The IV was feedback to students (verbal praise vs. no verbal praise). The DVs were as follows: DV 1, achievement; DV 2, homework preparation; DV 3, motivation to learn. There were three separate statistical analyses.

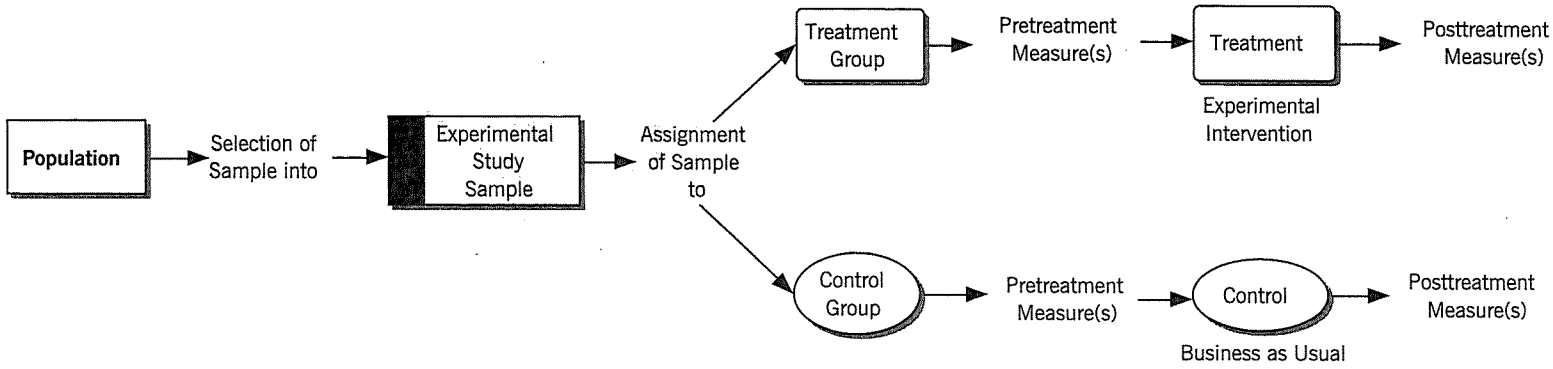


FIGURE 8.6. Basic pre- and posttest design.

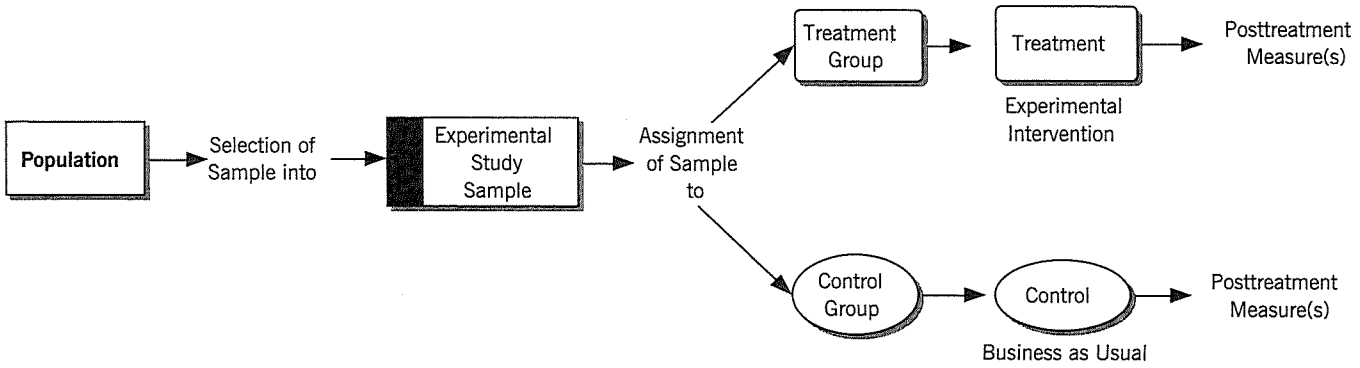


FIGURE 8.7. Basic posttest-only design.

Sampling

The sample consisted of 54 graduate students (49 females and 5 males) who were enrolled in two sections of a graduate-level course that the researcher taught. One intact class was designated as the control group; the other intact class was designated as the treatment group. This procedure for group assignment marks the study as a quasi-experiment; however, the matching of the subjects in the groups increased the likelihood that the groups were equivalent.

Data Collection

The measures for the DVs were as follows. Achievement was measured with an instructor-made examination that was vetted for validity and reliability; homework preparation was measured with a student-reported homework log of hours spent in homework; and motivation was measured with the Motivated Strategies for Learning Questionnaire (MSLQ), a standardized questionnaire that showed adequate validity and reliability. The measures were administered and scored at the end of the experiment.

Purpose and Design

For the purpose of investigating the effect of verbal praise on the three DVs, the researcher used a posttest-only design (see Figure 8.8). At five randomly selected class meetings, he viewed the homework logs of students in both groups, and he verbally praised subjects in the treatment group who had met the expectation of 180 minutes per week of home preparation. Subjects in the control group received a neutral response of "Thank you." During the last class, the instructor collected the homework logs, administered the MSLQ, and gave the final exam.

Data Analysis, Interpretation, and Conclusions

The *t*-test is often used with a randomized posttest-only design; here the researcher used three two-tailed *t*-tests, one for each DV. There were three steps involved in arriving at *p*-values: (1) calculation of the means and standard deviations for the posttest scores; (2) calculation of the *t*-values; and (3) consulting a table of critical *t*-values to see whether the calculated *t*-values met the alpha level the researcher had set.

Table 8.1 shows the results of the *t*-test analysis. The table presents the mean and standard deviation for control and treatment groups on each of the three DVs. The *t*-values are also indicated. The number in parentheses (52) before each *t*-value is the *degree of freedom* (often abbreviated as *df*, especially in equations), which is used to determine the *p*-value. The *df* is simply the number of subjects minus the number of groups (in this case, $54 - 2 = 52$). The asterisk (*) that appears next to each *t*-value indicates that $p \leq .05$ has been achieved; this is explained in the footnote at the bottom of the table.

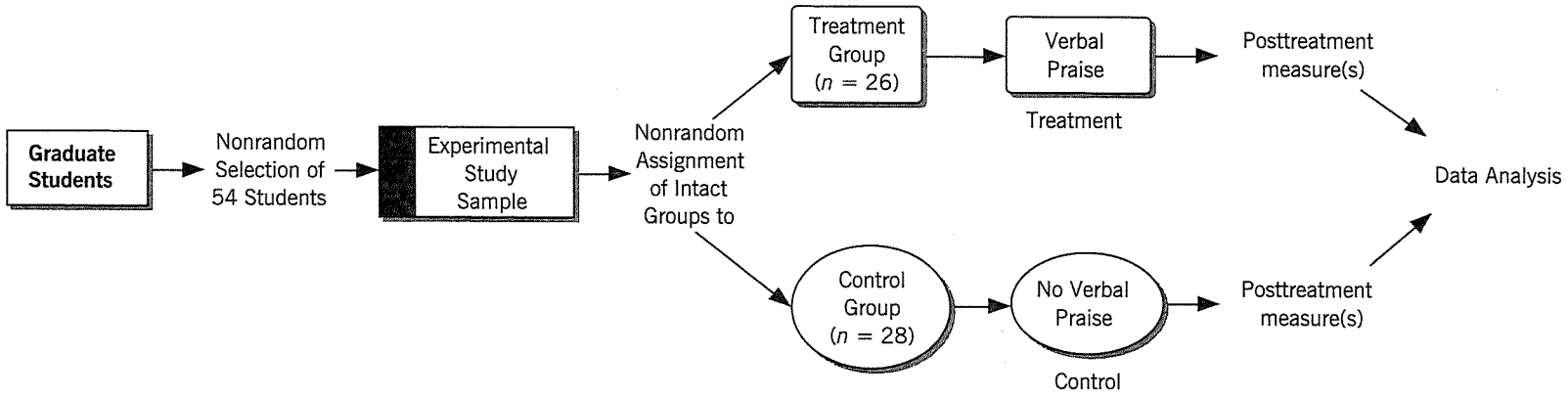


FIGURE 8.8. Design for the Hancock (2002) experiment.

TABLE 8.1. Results of the Hancock (2002) *t*-Test Analysis

	Teacher test		Homework hours		MSLQ	
	Control	Treatment	Control	Treatment	Control	Treatment
Mean	81.86	85.11	137.14	149.38	4.97	5.11
SD	6.06	5.49	15.17	22.16	0.251	0.221
	t (52) = 2.065*		t (52) = 2.38*		t (52) = 2.170*	

*$p \leq .05$.

In order to generate the p-value, the researcher consulted a table of critical t-values, illustrated in Table 8.2. At the top are the alpha levels for one- and two-tailed tests. The first column on the left lists degrees of freedom. The next six columns present the critical t-values for each type of test and at each alpha for each degree of freedom. Results are considered significant when the t-values from an experiment are higher than the critical t-value. Looking across the $df = 52$ row and down the column for a two-tailed test at alpha = .05, we can see that the critical t-value = 2.007 (underlined). Since the Hancock article reported higher t-values for differences in each of the three DVs, the three differences were judged to be statistically significant at $p \leq .05$. This also meant that the null hypotheses were rejected and the research hypotheses were confirmed.

In addition, the researcher reported effect sizes for the three DVs as follows: achievement, ES = 0.54; homework, ES = 0.80; motivation, ES = 0.56. In other words, the administration of verbal praise led to a one-half standard deviation increase

TABLE 8.2. Table of Critical *t*-Values for Hancock (2002)

	Alphas					
	Significance for a directional (one-tailed) test					
	.05	.025	.01	.005	.0025	.001
	Significance for a nondirectional (two-tailed) test					
df	.10	.05	.02	.01	.005	.002
50	1.676	2.009	2.403	2.678	2.937	3.261
51	1.675	2.008	2.402	2.676	2.934	3.258
52	1.675	<u>2.007</u>	2.400	2.674	2.932	3.254
53	1.674	2.006	2.399	2.672	2.929	3.251
54	1.674	2.005	2.397	2.670	2.927	3.248
55	1.673	2.004	2.396	2.668	2.925	3.245
60	1.671	2.000	2.390	2.660	2.915	2.232

Note. *df* = 50–60.

in achievement and in motivation, and a four-fifths standard deviation increase in hours of home preparation. These are substantial gains.

The researcher concluded that the three research hypotheses were confirmed, and that there were significant and ample differences in change on the DVs between the control and treatment groups.

Evaluation

The overall rating of this study is moderate, based on three criteria:

1. Theory and treatment were strong; construct validity was demonstrated; and the researcher controlled for threats to internal validity. The author begins the article with a review of more than 35 previous studies, to establish the theoretical foundation for the experiment. The IV and DVs were clearly defined, as were the three hypotheses.
2. Sampling was weak to moderate and threatened external validity. The author provides some detail about sample characteristics. The sample size was a bit smaller than 60 and was disproportionately female. Intact groups were assigned to control and treatment groups; however, the matching of the subjects in the groups increased the likelihood that the groups were equivalent.
3. Collection and analysis of data were strong, and statistical conclusion validity was demonstrated. The measures of achievement and motivation are clearly described, and the measures for achievement and motivation were vetted for validity and reliability. Analysis included significance/hypothesis testing (with an appropriate alpha) and effect size estimates. In describing results, the researcher stays close to the data and does not overconclude.

CHAPTER SUMMARY

MAP 8.7

✓ The purpose of an experiment is to investigate a predicted causal relationship when an intervention is introduced and manipulated by a researcher.

✓ The research review makes a case for the study and establishes a theory that guides the development of a hypothesis and the selection of variables.

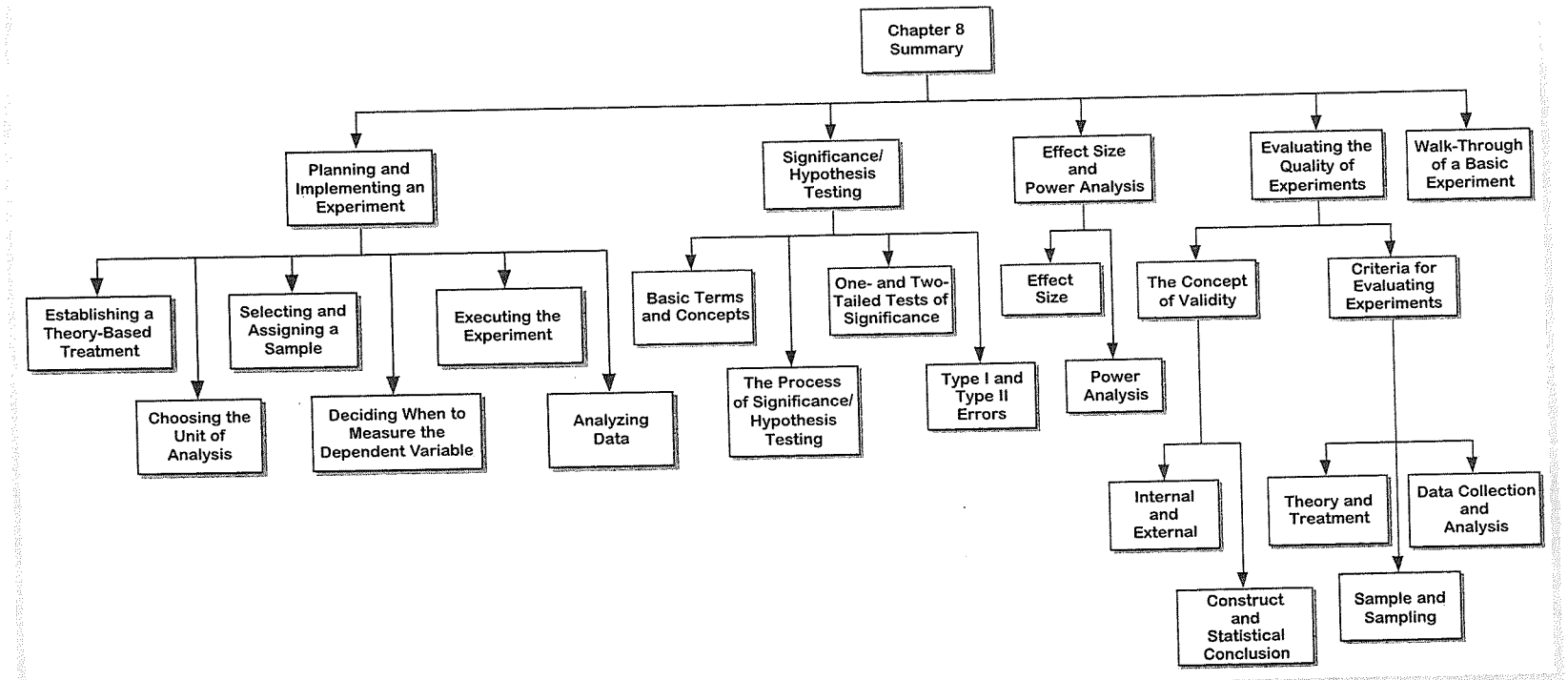
✓ Hypotheses may be directional or nondirectional.

✓ Independent variables (IVs) = causes; dependent variables (DVs)= effects.

✓ Researchers make the design decisions about (1) the unit of analysis (group or individuals); (2) sampling (size, selection, assignment); and (3) data collection (when and how often to measure the DV).

✓ Between-group designs have a control group and a treatment group.

✓ Experiments use randomized samples from a population or convenience samples; sample size should be at least 60.


Chapter 8 Summary
Planning and Implementing an Experiment
Establishing a Theory-Based Treatment
Choosing the Unit of Analysis
Selecting and Assigning a Sample
Deciding When to Measure the Dependent Variable
Executing the Experiment
Analyzing Data
Significance/ Hypothesis Testing
Basic Terms and Concepts
One- and Two-Tailed Tests of Significance
The Process of Significance/ Hypothesis Testing
Type I and Type II Errors
Effect Size and Power Analysis
Effect Size
Power Analysis
Evaluating the Quality of Experiments
The Concept of Validity
Internal and External
Construct and Statistical Conclusion
Criteria for Evaluating Experiments
Theory and Treatment
Data Collection and Analysis
Sample and Sampling
Walk-Through of a Basic Experiment

MAP 8.7

- ✓ True experiments randomly assign subjects to groups; quasi-experiments do not use randomized group assignment.
- ✓ Analysis of data always involves significance testing and hypothesis testing; they may also include effect size estimates and power analysis.
- ✓ Statistical significance means a low probability (p) that a difference between groups is due to sampling error, and a high likelihood that it is due to the treatment.
- ✓ The t-test is the simplest way to establish the statistical significance of the difference in results between one control and one treatment group.
- ✓ $p \leq .05$ is the threshold (alpha or α) conventionally used for statistical significance.
- ✓ Inferential/significance tests are based on the normal (bell) curve and may be one- or two-tailed.
- ✓ Hypothesis testing uses alpha (α) to determine the probability that the null hypothesis is false.
- ✓ Type I error, also known as the alpha-level error, occurs when the researcher rejects the null hypothesis even though this hypothesis is true; Type II error, also known as the beta-level error, occurs when the researcher accepts the null hypothesis even though this hypothesis is false.
- ✓ Effect size (ES) is a standardized number that describes the overall strength of an experimental intervention; ES = 0.5 is considered adequate.
- ✓ Statistical power tells the likelihood that the researcher has established an adequate sample size and will thus avoid making a Type II error. Power = .80 is desired.
- ✓ Validity of a study refers to the strength of the inferences and conclusions that can be drawn, and the level of confidence in results.
- ✓ Internal validity answers this question: Are the changes in the DV due to the treatment, or are there alternative explanations (threats to internal validity) that might account for the differences?
- ✓ External validity answers this question: Can the findings be generalized to other individuals, settings, and times? There are two types of external validity: population validity and ecological validity.
- ✓ Construct or theoretical validity answers this question: How well does the research review support the operational definitions used in the research question and hypotheses?
- ✓ Statistical conclusion validity answers this question: Do the measurement and statistical argument support the relationship between the IV and DV?
- ✓ Threats to validity are alternative explanations that can account for results.
- ✓ Evaluating an experiment depends on three sets of criteria: theory and treatment, sample and sampling, and collection and analysis of data.

KEY TERMS AND CONCEPTS

alpha
construct/theoretical validity
control group
degree of freedom (*df*)
dependent variable (DV)
directional hypothesis
ecological validity
effect size (ES)
external validity
hypothesis testing
independent variable (IV)
internal validity
nondirectional hypothesis
null hypothesis
one-tailed test
population validity
power/power analysis
p-value
quasi-experiment
research hypothesis
statistical conclusion validity
statistical significance
table of critical *t*-values
threats to validity
treatment group
true experiment
two-tailed test
Type I (alpha-level) error
Type II (beta-level) error
validity

REVIEW, CONSOLIDATION, AND EXTENSION OF KNOWLEDGE

1. Match the terms in Column A to the definitions in Column B.

Column A	Column B
independent variable	magnitude of difference
true experiment	the causal variable
alpha	rejects the null hypothesis when it is true
null hypothesis	effect is due to the cause
effect size	random assignment to groups
Type I error	test for significance
Internal validity	$p \leq .05$, used in hypothesis testing
t-test	predicts no difference in outcomes

2. Using an electronic database, search for an article describing an experiment on a topic of interest to you. Follow the Guide to Reading below as you work through the article. Then, with the Guide as a template, write a critique of the article you have selected.

Guide to Reading

Research Review: Does the review establish a theoretical foundation for the experiment? How current is the review? Are more than 10 studies reviewed? What hypothesis/hypotheses or research questions are generated from the research review?

Purpose and Design: What was the purpose of the study? What were the IV and DV(s)? How were they operationalized? What was the design of the study?

Sampling: What was the sample size? How was the sample selected (randomly or nonrandomly)? What were the characteristics of the sample? How was the sample assigned to groups (randomly or nonrandomly)?

Data Collection: What specific data collection strategies were used? Do the authors give indications of the measures' validity and reliability? If so, what are they? How well did the data collection strategies match the research question or purpose?

Data Analysis, Interpretation, and Conclusions: How were the data analyzed? What statistical tests were used? What were the results? Were the results significant or nonsignificant? How well does the data analysis match the research question or purpose? To what degree do the analysis, interpretation, and conclusions stay close to the data? Have the researchers overconcluded?

Evaluation: How do you rate the overall quality of the study? Strong? Moderate? Weak? How do you rate the categories below, and what is your reason for each rating?

Theory and treatment: Strong, moderate, weak

Sample and sampling: Strong, moderate, weak

Data collection and analysis: Strong, moderate, weak

Sample Critique

This sample critique is a review of the same article covered in the "Walk-Through of a Basic Experiment" section of this chapter, with more details included.

Hancock, D. R. (2002). Influencing graduate students' classroom achievement, homework habits and motivation to learn with verbal praise. *Educational Research, 44*(1), 83–95.

Research Review

The article begins with a review of more than 35 previous studies to establish a theoretical framework for the study. The articles reviewed focused on the relationship between verbal praise and "motivation to learn, classroom achievement, homework habits, and graduate education" (p. 83), as well as developing a detailed description

of well-constructed verbal praise. The researcher developed three hypotheses from the review (pp. 88–89):

1. Postsecondary students exposed to well-administered verbal praise by a professor would demonstrate higher achievement levels on a professor-made examination than would students who received no well-administered verbal praise.
2. Postsecondary students exposed to well-administered verbal praise by a professor would spend significantly more time preparing at home (i.e., doing homework) for each lesson than would students who received no well-administered verbal praise.
3. Postsecondary students exposed to well-administered verbal praise by a professor would demonstrate higher motivation levels to learn in the classroom than would students who received no well-administered verbal praise.

Purpose and Design

The purpose of the study was to investigate the effect of well-constructed verbal praise on student outcomes. The IV was well-constructed verbal praise. The DVs were as follows:

DV 1 = achievement

DV 2 = homework preparation

DV 3 = motivation to learn, measured by the Motivated Strategies for Learning Questionnaire (MSLQ)

Sampling

The convenience sample consisted of "54 graduate students in a one-semester course in Educational Research Methods that the researcher taught at a middle-size[d], state-supported university in the southeastern United States" (p. 86). The 49 female and 5 male students were enrolled in one of two relatively equal-sized sections. The researcher designated one section as the treatment group and the other as the control group. This intact groups sampling strategy is indicative of a quasi-experiment. The researcher noted that the sections were matched with respect to age, gender, degree program, GPA, and socioeconomic status.

Data Collection

The measure of home preparation was a self-reported homework log. The measure for motivations was a portion of the MSLQ, a 7-point Likert scale questionnaire that was administered during the 15th class session. The MSLQ assesses motivational tendencies, self-efficacy, and success expectancy. Its predictive validity was reported as .29; internal consistency ranged from .62 to .93. The measure of achievement was an

instructor-created final examination administered during the final classes. Content validity was established by having two experts review the test's alignment to lesson objectives; the split-half reliability coefficient was .76.

Students were told that there was an expectation for homework of about 180 minutes weekly. The instructor viewed the logs during five randomly selected classes. In the treatment group, the instructor verbally praised students who had met the expectations. Students in the control group received a neutral response of "Thank you." An independent observer checked the fidelity of the treatment.

Data Analysis, Interpretation, and Conclusions

Data were analyzed with three *t*-tests, one for each measure. Results for the three DVs were as follows: for achievement, $t = 2.065$; for homework, $t = 2.38$; for motivation (MSLQ), $t = 2.170$. Each finding was significant at $p \leq .05$. The results indicated that the effects were most likely due to the treatment and not to chance or error. The researcher rejected the null hypotheses and accepted the research hypotheses as listed. In addition, the researcher conduced an effect size estimate that yielded these results: ES = 0.54 for achievement; ES = 0.80 for homework; and ES = 0.56 for motivation.

Evaluation

Using the criteria of theory and treatment, sample and sampling, and data collection and analysis, I give this study an overall rating of moderate.

- *Theory and treatment* merit a rating of strong. The article begins with a review of more than 35 previous articles to establish the theoretical foundation for the experiment. The researcher clearly defined the IV and DVs at the start of the research.
- *Sample and sampling* merit a rating of only weak to moderate. The author has provided a description of the sample size and some details about sample characteristics. However, the sample size was below 60; the sample was disproportionately female; and there was nonrandom group assignment. As a result, the sample does not lead to the generalization of findings.
- *Data collection and analysis* are rated as strong. The measures of achievement and motivation are clearly described, and the measures for achievement and motivation were vetted for validity and reliability. Hypothesis testing may be assumed from the alpha level and *t*-values. The researcher stays close to the data and does not overconclude.