Chapter 1

Introduction

The purpose of this book is to train students—potential researchers and consumers of research—to critically read a research article from start to finish. You will learn to critically read an introduction, which sets the stage by describing the rationale for the study (i.e., what led to it) as well as its purpose (i.e., what the study hoped to accomplish). You will learn how to "dissect" the method section so that you can decide whether precautions were taken to guard against threats to internal validity, both in terms of assignment of participants to the various conditions of the study and in use of control procedures and groups. You will become more familiar with interpreting results and even with performing additional calculations or checking a particular result. Finally, you will learn to carefully evaluate the experimenter's discussion of the results to determine the extent to which the conclusion is justified, can be generalized, and has limitations.

Studies are presented in order of increasing complexity. Each is prefaced with an introduction that describes the basic design and statistical analysis. Every effort has been made to locate examples of good as well as flawed studies in each of the categories, ones that performed statistical analyses that are commonly taught at intermediate and advanced levels. When it is feasible, articles are presented verbatim. For the most part, however, sections have been excerpted and/or revised for clarity. Ellipses and bracketed phrases are used to indicate some changes. More extensive revisions or synopses are indented from the margins, enclosed in brackets, and always appear in italics.

All of the chapters (with the exception of Chapter 3) contain two examples of studies that employed a particular design. You will begin by evaluating the first study with our assistance. Copious

notes are added to alert you about potential flaws or positive aspects of the design. These are indicated by an arrow in the top left; they are enclosed in parentheses and always appear in italics. When we evaluate articles together, questions and answers are shaded to make them distinct. The second article is excerpted, and it is followed by a series of critical guide questions. Following questions, a page number directs you to the answers, which are found in a separate section at the end of the book.

The remainder of this chapter presents a review of the bases of empirical research (controlled observations that are reliable and valid) followed by a review of potential trouble spots that can invalidate a conclusion about the effectiveness of an independent variable.

A SCIENTIFIC APPROACH

Before discussing the nuts and bolts of empirical research, let's discuss the scientific endeavor more generally, in order to place the methods discussed in this book within a useful context. Much of the philosophy of science is concerned with the nature of cause and effect. Philosophers from a wide range of traditions (positivist, essentialist, activity theorists, and evolutionary critical realists) have all been concerned with the following question: "How can we determine the cause of an event?"

Research in the social sciences has been strongly influenced by the writings of John Stuart Mill and Karl Popper. Mill held that causal inference requires that (a) cause has to precede effect in time, (b) cause and effect must be related, and (c) all other explanations for the relationship must be eliminated. If the presumed cause is present whenever the effect is present (method of agreement), the effect is absent whenever the cause is absent (method of difference), and both can be observed repeatedly (method of concomitant variation), we have evidence for causality. For Popper, the most important feature of the scientific approach involved the falsification (rather than the confirmation) of theories. Differing explanations (theories) for observations are placed in competition with each other. The theory that best explains the data (in terms of simplicity, predictive power, and the ability to incorporate new data) is retained, until replaced by a better theory.

Both Mill and Popper had a healthy appreciation for the need to examine and eliminate alternative explanations for the findings before settling on a presumed cause. In the social sciences, the removal of such alternate explanations (called confounding factors) is a constant struggle, as you will see throughout this book.

The scientific method typically follows the hypothetico-deductive approach outlined here:

1. Make observations about a phenomenon.

2. Form hypotheses (proposed explanations for the observed phenomenon).

3. Make predictions based on these hypotheses.
4. Test the predictions through observation and experimentation.
5. Based on the results, form new hypotheses (Step 2), and repeat Steps 3, 4, and 5 in an iterative fashion.

To the degree that a theory leads to testable predictions, which are confirmed in repeated assessments by multiple scientists in varying settings and cannot be explained by confounding factors or better alternate theories, the greater weight and stature it is given.

EMPIRICAL RESEARCH

There are three key concepts associated with empirical research: controlled observation, reliability, and validity. **Controlled observation** refers to the precision of conditions under which data are collected. In essence, any "noise" that can affect the data is eliminated, minimized, or counteracted in such a way that any other observer can replicate the conditions. Superfluous factors in the environment are eliminated or minimized by gathering data under uniform conditions; environmental distractions (e.g., sights or sounds) are the same for all participants, surrounding temperature is the same, and the data collector is the "same" (at the same level of expertise throughout testing)—or if more than one is used, they are equally distributed throughout the various groups and so forth. These features ensure that the collected data will be objective, precise, and verifiable.

Reliability refers to a broad range of phenomena. Reliability means repeatability or consistency. Empirical research should be reliable: Under the same experimental conditions, anyone else should be able to obtain the same results (i.e., the outcome of data collection should lead to the same conclusion). Reliability also refers to precision of our measuring instruments. Precise instruments are more likely to yield consistent measures than cruder instruments. For example, a determination of 4 oz (113.398 g) of liquid will be more reliable if a calibrated measuring cup rather than a drinking glass is used.

Reliability also refers to the extent to which a test measures consistently or yields a "true" or accurate measure of whatever it is the test measures. That accuracy shows up in two ways: the repeatability of the score on more than one occasion and the same relative standing of the individuals in their group on more than one occasion.

Validity is the final key concept of empirical research. It is synonymous with appropriateness, meaningfulness, and usefulness. With regard to research, we want to know whether conclusions are valid; are they appropriate, meaningful, and useful on the basis of the intent of the investigator and the procedures used to fulfill that intent? There are three types of study validity that we must be concerned with: internal validity, statistical conclusion validity, and external validity.

A study has **internal validity** to the degree that it allows us to conclude that a relationship between variables is causal *or* that the absence of a relationship implies a lack of cause. **Statistical conclusion validity** refers to the appropriateness of the statistical methods employed to determine if covariation exists or not. A study has **external validity** to the degree that the results can be generalized beyond the current study to situations that use other measures, methods, and populations. Our goal in research is to devise studies that allow us to derive clear and unambiguous answers to the questions posed. To do this, we seek to design our research to limit, to the greatest extent possible, the threats posed to each of form of validity.

THREATS TO INTERNAL VALIDITY

To evaluate the soundness of each design, you need to keep in mind potential sources of **confounds** (other potential explanations of results). These are variables that may be operating in conjunction with the manipulated independent variable and make it impossible to determine whether observed changes or differences in the dependent variable are due to the manipulation, the confound, or a combination of the two. Because these potential confounds may threaten the extent to which the conclusion is valid or justified (i.e., internal validity of the study), they are called threats to internal validity. Some threats to internal validity apply to study designs that incorporate pretests and posttests. Some threats apply to research designs in general.

■ *Studies With Pretests and Posttests*

History. This refers to any event occurring in the interim that directly or indirectly could affect the behavior being measured and therefore also could account for the results.

Initial testing. This refers to a change in posttest performance that results from pretest experience.

Instrumentation. This refers to any change in the measuring instrument and/or assessor from pretest to posttest that can just as easily explain a change in scores.

Maturation. This refers to any change within the participant that occurs during the interim and can just as easily account for posttest performance.

Regression toward the mean. This is a predicted shift in posttest scores when participants were specifically selected because their pretest scores were extremely high or low. Posttest scores are predicted to be less extreme, regardless of treatment effects.

■ *Research Situations (With or Without Pretests and Posttests)*

Compensatory equalization. This refers to the administration of some treatment to a control group to compensate for its lack of the beneficial

treatment being received by an experimental group. This reduces differences between posttreatment means of the groups.

Compensatory rivalry. This refers to behavior of a control group such that participants attempt to exceed performance of an experimental group because they are not receiving equal treatment. This reduces posttreatment differences between groups.

Diffusion of treatment. This is the unintentional spread of treatment to a control group (or groups) when participants receive information withheld from them (e.g., through conversation with experimental participants) that results in a smaller difference among group performances at posttreatment assessment.

Experimenter expectancy. This refers to a characteristic of the individual who is collecting the data. When a researcher (e.g., author of the article) tests the participants, his or her expectations for certain results unintentionally may affect participants so that they behave in accordance with the hypotheses. Concomitantly, recording errors may be made, also in the direction of a hypothesis.

Hawthorne effect. This refers to a change (usually positive) in participants' behavior because they were assigned to a treatment group rather than because of the treatment itself.

Interaction effects. These refer to threats that operate on a select group of individuals that also could account for observed results (e.g., a historic event that affects one particular group of participants).

Resentful demoralization. This is a lowered level of performance by a control group because participants resent the lack of experimental treatment. This increases the differences between posttreatment group means.

Selection bias. This refers to the assignment of participants to the various test conditions on a nonrandom basis. Differences in performance may be associated with a participant characteristic instead of, or along with, the independent variable.

Selective loss (mortality, attrition)**.** This is the loss of particular participants from a group (or groups) in such a way that remaining participants no longer can be considered to be initially equivalent with respect to the dependent variable.

THREATS TO STATISTICAL CONCLUSION VALIDITY

In most instances of experimentation, the conclusions reached by the researcher are based on the outcomes of statistical analyses. Typically, this involves rejection or retention of a null hypothesis. When the null hypothesis is retained, the researcher concludes that there was insufficient evidence for a difference between group means (or whatever statistic was being evaluated). If treatment truly was ineffective,

the conclusion is correct. However, if treatment effectiveness simply was not evident in the statistical analysis, the conclusion is erroneous, a **Type II error**. When the null hypothesis is rejected, the researcher concludes that there was evidence for a difference between group means and that the independent variable was effective. If this is so, the conclusion is correct. However, if treatment actually was ineffective, then the researcher's conclusion is erroneous. A **Type I error** has been committed. Box 1.1 summarizes the Type I and Type II errors.

Box 1.1 Summary of Type I and Type II Errors

Type I: Rejecting the null hypothesis when it is true

Type II: Retaining the null hypothesis when it is not true

Any factor that leads to a Type I or Type II error is a threat to validity of the statistical conclusion. Briefly, these threats include

Fishing. When an unreasonable number of statistical tests are conducted on the same data, one may reveal what *appears* to be a significant difference by chance alone. This threat, which leads to a Type I error, can be reduced by statistical adjustments that effectively lower the probability level needed to declare any comparison significant.

Insufficient power of statistical test. This is one of the most prevalent causes of a Type II error. Power refers to the likelihood of rejecting a null hypothesis that is false or correctly declaring a difference in some statistic significant. Assuming that a difference of a particular magnitude is anticipated, sample sizes have to be large enough to detect that difference. Often, because of research constraints, an insufficient number of participants are tested, not enough to reveal the effect of an independent variable, particularly one whose effect is of small magnitude. Had the sample been large enough, had the significance level of the statistical test been less stringent, and/or had a more powerful statistical test been performed on the data, the effect would have been detected. Better studies include a prior analysis of the sample size required to identify a difference or effect of a given magnitude at a given significance level when estimates of variance (variability in performance) are available. However, the opposite should be mentioned. If samples are extremely large, almost any effect can be significant, even if it is so small as to be unimportant from a practical standpoint.

Unreliable instrument. If the measuring instrument is not reliable (does not measure consistently), performances will be variable, the error term of the statistical test will be inflated, and a true difference between means may not be evident.

Varied participant characteristics. If participants differ in age, gender, intelligence, or other characteristics that are related to the dependent variable measure, performances will be variable, and a true difference between means may not be evident.

Varied test conditions. If testing conditions are not uniform, within and between groups, performances will be variable and a true difference between means may not be evident.

Violation of statistical assumptions. All statistical tests have underlying assumptions. If at least one is seriously violated and the statistical test is not robust with respect to violation of certain assumptions, the analyses may fail to reveal a difference that truly exists (Type II error) or may reveal a difference that is really due to chance (Type I error).

THREATS TO EXTERNAL VALIDITY

Threats to external validity include any factors that could limit the generalizability of the study findings. Factors that limit generalizability include idiosyncratic features of the sample used and specific aspects of the study context (methods, procedures, and measures). For example, studies of psychotherapy treatment typically involve patients who have volunteered to participate. Would the results regarding treatment efficacy generalize to patients who had not volunteered? If the therapists were primarily young and female, would the results generalize to situations in which the therapists were older and male? If depression improvement was defined by decreased scores on the Beck Depression Inventory (Beck, Steer, & Brown, 1996), would similar results be obtained using a different self-report measure or observer ratings? If the study was conducted in a New York metropolitan hospital, would the same results be expected in a rural Minnesota clinic? To the degree that idiosyncratic features of the study limit generalizability, we have threats to external validity.

MEASUREMENT AND PSYCHOMETRICS

The quality and characteristics of any instruments used to measure the independent and dependent variables are intimately connected to a study's reliability and validity. In this section, we will briefly review the assessment of reliability and validity of psychological instruments.

■ *Reliability of Measurement*

Measurements have two parts to them, although in practice it is hard to separate the parts. One is called **systematic variance**, the accurate measure of the characteristic, and the other is nonsystematic or **error variance**, a part of the score that is due to factors other than the characteristic (e.g., the instrument was crude, the person doing the measuring was erratic, the person's motivation changed). As error variance decreases, test reliability increases.

Several procedures can be used to establish reliability; one is **test–retest reliability**. The same group of people is given the test on two occasions. This measures

stability of the scores. If the scores each time really measure the characteristic accurately, the same scores should approximately be achieved each time, and the individuals' relative standing in the group should be the same (the highest score should be achieved by the same person each time, etc.). This applies to tests that measure relatively stable characteristics (e.g., intelligence) as opposed to unstable characteristics (e.g., mood). Instead of using the same test on two occasions, one can use alternate forms of the test, if available, which yield alternate-form reliability, or **equivalence reliability**.

Another procedure is to compare scores within one test administration to measure **internal consistency**, the extent to which items within the test measure consistently. One may compare scores on one half of the test with those of the other half. This is interitem or **split-half reliability**. You could literally look at the first half versus the second half or even numbers versus odd items. The rationale is that if all items measure the same characteristic, there should be a relationship, or correlation, between scores achieved on each half of the test.

The **Kuder-Richardson Formula 20** is used to measure reliability of a test whose items can be scored right or wrong, and it measures the extent to which all participants answered each item appropriately. To the extent that they did, the items are considered equivalent, and the test is considered reliable. Similarly, the **Cronbach's alpha** reflects the extent to which there is agreement among participant responders on a test whose items are scaled or weighted. An example would be a Likert-like test in which numerical response choices might reflect ranges from *strongly agree* to *strongly disagree*. In all instances, measures of reliability are in the form of a **correlation coefficient** that ranges from 0 to 1.

There is an additional form of reliability that can be determined. If participants have to be judged on the extent to which they display some behavior, it is desirable that more than one judge does the rating and that we have some basis for deciding how consistent the judgments are. This is a problem approached by measuring **interrater reliability**, if a correlation is determined, or **interrater agreement**, if you are looking for percentage agreement. As a general rule of thumb, coefficients should be at least .65 for any measure to be considered reliable.

■ *Validity of Measurement*

Validity also is a central issue in measurement and testing. Tests are valid to the extent that they measure the characteristics they were intended to measure. **Content validity** refers to the extent to which the items reflect an adequate sampling of the characteristic. For instance, do items of a mood test sample only positive moods, or do they sample positive, negative, sad, and apprehensive moods? **Criterion validity** refers to the extent to which test scores correlate with a behavior (the criterion) the test supposedly measures (concurrent validity) or the extent

to which test scores predict that behavior (predictive validity). For example, scores on an intelligence test should correlate with scores on an achievement test and should predict college performance.

Finally, **construct validity** refers to the extent to which the test measures the characteristic it intends to measure, because the characteristic cannot be measured directly. Of the many types of validity, construct validity is most difficult to establish. It is typically approached through a process of building up convergent and discriminant validity evidence. If scores on the test correlate with other measures of that characteristic, this would yield convergent validity. For example, scores on a trait anxiety scale should correlate with behavioral measures in anxiety-provoking situations. If scores on a test do not correlate with measures of other characteristics, using the same test or other tests or measurements, this would be evidence of discriminant validity. For instance, scores on an intelligence test should not reflect introversion or extroversion and should be unrelated to scores on a test of self-esteem or mood. To the extent that an experimental procedure varies the intended construct (characteristic), all other things being equal, statements about it will be valid—that is, appropriate, meaningful, and useful.

PLAN OF THE CRITIQUES

Questions will appear throughout the articles we are evaluating together. They are designed to guide you step-by-step as you read and are questions that you should be asking yourself while you are reading an article. Two parts of an article will not be addressed: the abstract—a concise summary of the study—and references. You will start with the rationale for the study. This lets you know what, in the past literature, aroused the investigator's interest in the subject matter to begin with: inconsistent or contradictory findings, a possible confound in earlier studies, curiosity, a logical deduction from a theory that could be tested, and so forth. Always ask yourself: What was the rationale for the study?

Next, you move on to the purpose of the study. This lets you know exactly what the investigators intended to accomplish. Often, it is expressed in terms of testing particular hypotheses. Sometimes it is expressed in terms of what the authors wanted to demonstrate or determine. Always ask yourself: What was the purpose, or reason, for conducting this study?

The method section is next. You will want to focus on participants first. Who were tested, and how were they recruited? Were they randomly selected or assigned, matched, lost because of attrition? Here is where you want to consider that groups may not have been initially equivalent. Next, you might want to look at tests that were used. If they are not well known, you want assurance that they are reliable and valid, and always consider the possibility that they might not be so. If more

than one is included, you want assurance that they were presented in counterbalanced order. Finally, the rationales given for using the particular tests should indicate that they are appropriate for fulfilling the purpose of the study.

The procedure section focuses on what was done. A question about general procedure assesses your basic understanding of what was done to the participants or what they were required to do. Specific questions alert you to possible sources of confounds (e.g., a shift in testing conditions, a failure to assess a manipulation, testing performed by the researcher rather than by a naive experimenter).

The questions relating to results of the study focus on appropriate analyses of the data. Readers typically assume that no mistakes have been made and that assumptions underlying all tests have been met. This is not always true, and the outcome can be serious: A mistake in calculations can change the conclusion reached by the investigators. Hopefully, you will remember to check on the accuracy of degrees of freedom for independent and dependent *t* tests and **analysis of variance (ANOVA)**, as well as appropriateness of the alpha levels used, especially for planned and post hoc comparisons. These points will be reviewed in the introduction to the article. Specific questions address these issues. Consider them as quick checks. If they are correct, hopefully, the calculations are accurate.

Unfortunately, this will not tell you whether the statistical test is appropriate or if basic assumptions have been met. But even here quick checks are possible. If standard deviations are given, you can square the values to obtain variances and form simple ratios of the largest to the smallest to see whether there is homogeneity of variance. If the study involves repeated measures, you can check the significance of F ratios with $df = 1$ and $N - 1$ to see whether they are still significant. This checks the validity of statistical conclusions if lack of circularity—equality of variances of differences—has not been considered. If F ratios still are significant, results (at least statistical) are valid. If not, conclusions about means differences may not be warranted. Again, each will be reviewed with the relevant article.

The final section of the report is the discussion. It is here that the researchers reach some conclusion regarding the outcome of their manipulation—for instance, its effectiveness. It is here that questions deal with the validity of the conclusion. You will be asked to consider that threats to internal validity might not have been eliminated, rendering the conclusion unjustified. And because the intent of the study is to generalize the results, you may be asked to consider issues related to external validity (i.e., limitations in the extent to which results will generalize).

The relevant features of a critique are summarized in Box 1.2.

Box 1.2 Essential Features of a Critique

Feature	Relevant Question(s)
Rationale	What is the reason for conducting the study?
Purpose	What does the researcher intend to accomplish?
Method	
Participants	How were participants selected? Assigned? Which, if any, were lost?
Apparatus	Were measuring instruments reliable? Valid?
Procedure	What did participants do? How were they measured? With what? Who measured them? Were testing conditions uniform? Were factors other than the independent variable operating?
Results	How were data analyzed? Was analysis appropriate? Accurate?
Discussion	What were major conclusions? Are they justified (valid)? Can results be generalized? To which population(s)?

Before beginning our detailed analyses, let's consider two studies in abbreviated form—both of which compare two group means but differ in the extent to which valid conclusions can be drawn. Both use two distinct groups of participants. In both cases, the groups are not formed by random assignment. Instead, they are based on differences that exist between the participants of each group. The first study does not attempt to match the participants on important variables that could account for group differences just as easily as the factor that makes each group unique. The second study does attempt such a matching procedure.

STUDY EXAMPLE 1.1: "A COMPARISON OF OLDER AFRICAN AMERICAN WOMEN ABOVE AND BELOW POVERTY LEVEL"

The present study compared reported health-promoting lifestyles among older African American women who lived below or above the poverty level of $7,360. Thus, poverty level is the basis for differentiating the two groups.

The Study

Brady, B., & Nies, M. A. (1999). A comparison of older African American women above and below poverty level. *Journal of Holistic Nursing*, *77*(2), 197–207.

Poverty has been a major barrier to a healthy lifestyle among African American women in terms of their health status, use of health services, and mortality. The elderly have chronic health problems that are attributed to obesity because of lack of exercise, a sedentary lifestyle, and earlier age at first childbirth. Research suggests that exercise can reduce the risk of some of these chronic health problems by enhancing the quality of life of the elderly. Two thirds of African American women (vs. more than one half of all women) do not exercise. There are few studies of the health-promoting benefits for older African American women. Moreover, studies show that income level is correlated with health behavior, but this association in African Americans has not received much attention and has not been related to exercise. The purpose of this study was to compare the health-promoting lifestyles and exercise behaviors of African American women, above and below the poverty level, who lived in a community. The hypothesis was "that older African American women living above the poverty level will practice more health-promoting behaviors as measured by the Health-Promoting Lifestyle Profile (HPLP) than women living below the poverty level."

Method

Design, Sample, and Setting

A descriptive study design was used for this pilot study. The convenience sample consisted of 58 African American women, 50 years of age and older, living in the community. The participants were recruited with the assistance of the pastor from a local Baptist church in the mid-South. All women 50 years of age and older attending the church were invited to participate in the study. A notice was placed in the church bulletin [in order] to encourage participation. The data collection took place on a Sunday after service in the Baptist church Bible study room. A brief description of the study was given to the women by the researcher and informed consent was obtained before completing the instruments.

▶ *(Note that all participants were volunteers who were asked to participate in the church after church services. Thus, some may have been more willing than others to take part—additionally all were Baptists.)*

Instruments

Instruments . . . included a demographic data sheet and the HPLP. . . . The demographic data sheet included information about (a) age, (b) income, (c) marital status, (d) education, and (e) assistance with income.

The HPLP is a 48-item measure with a Likert response format: *never, sometimes, often,* or *routinely.* . . . The six HPLP subscales and sample items are as follows: *self-actualization* ("am enthusiastic and optimistic about life"), *health responsibility* ("have my blood pressure checked and know what it is"), *exercise* ("engage in recreational physical activities"), *nutrition* ("eat breakfast"), *interpersonal support* ("discuss personal problems and concerns with person close to me"), and *stress management* ("take some time for relaxation each day").

On development, the HPLP was found to have high internal consistency, with an alpha coefficient of .92. The six subscales were also found to have a relatively high internal consistency. The correlations for the scales ranged from .90, for self-actualization, to .70 for stress management . . . Test–retest reliability was $r = .93$ for the total scale and ranged from .81 to .91 for the subscales. . . . The instrument has been used with various ethnic groups, including . . . African Americans . . . and provided reliable and valid data.

The HPLP was scored by obtaining the mean of all 48 items on the total scale. The subscale of exercise was scored in a similar manner. Possible scores of the HPLP and exercise subscales ranged from 1.0 to 4.0. The alpha reliability coefficient of the HPLP in this study was .90. The exercise subscale of the HPLP consisted of five self-reported exercise behaviors. The alpha coefficient for the exercise subscale was .69.

▶ *(Note that the test was shown to be reliable, but validity measures are not reported. Furthermore, the sample question about nutrition asks about eating breakfast, which may or may not include a nutritious one. Likewise, the exercise subscale asks about "recreational physical activities," yet such activities as walking to a bus stop and housecleaning also are physical activities. Notice, too, that the reliability coefficient for this scale was lower than that for the group on which reliability first was established.)*

Procedure

. . . Women entered the room, and each was asked to sign a consent form before being handed the demographic sheet and HPLP instrument. . . . A brief description of the study and the importance of the study were given by the researcher.

The importance of filling out the instruments and the availability of the researcher to assist if needed were discussed. An African American female facilitator assisted with data collection and encouraged women to participate in the study. On completion of the instruments, each was collected and examined by the researcher for completeness. . . . [Data were organized] for participants living above and below poverty levels.

▶ *(Note again that the women were encouraged to participate. Moreover, the researcher was available to assist anyone needing help in filling out the questionnaire and, if anyone did require assistance, inadvertently may have influenced the way some of the questions were answered. Most important, the two groups were formed on the basis of information provided in the demographic sheet. Although this was the only way to form the groups, poverty level is not the only factor that differentiates them. A more desirable procedure would be to match the two groups on other relevant variables, such as education and health status.)*

Results

. . . The mean age of the 58 African American women was 60.1 years, with a range of 51 to 88 years. Table 1.1 shows that of the 58 African American women, 26 (44.8%) had individual incomes below the poverty level and 32 (55.2%) had individual incomes above the poverty level of $7,360. Their education ranged from less than high school to college graduates. Frequency distribution of marital status and assistance with income (financial support provided by another person) are also shown in Table 1.1.

▶ *(Note that there is no way of telling from Table 1.1 what percentage of those above poverty level had graduated from high school and college. Presumably, this would apply to a large percentage, but it is not necessarily so. Likewise, a large percentage were married or widowed, but we don't know their economic level. Married women, for example, would be more likely to receive interpersonal support because their spouses are readily available. If they were mainly among the above-poverty group, this in part could explain their higher scores on the HPLP.)*

The research hypothesis was supported. Participants living above the poverty level had higher overall scores on the HPLP ($M = 2.85$, $SD = .40$) than African American women living below the poverty level ($M = 2.51$, $SD = .51$), $t(56) = -2.79$, $p = .007$.

Table 1.1 Demographic Characteristics ($N = 58$)

Characteristics	*Group*	*n*	*%*
Socioeconomic status	Above poverty	32	55.2
	Below poverty	26	44.8
Assistance with income	None	32	55.2
	Family	7	12.1
	Spouse	13	22.4
	Friend	2	3.4
	Missing	4	6.9
Education	Below high school	12	20.7
	High school	30	51.7
	Some college	10	17.2
	College graduate	5	8.6
	Missing	1	1.7
Marital status	Married	19	32.8
	Divorced	6	10.3
	Single	5	8.6
	Widowed	28	48.3

The range of scores of the total HPLP, along with the range of scores for the two income groups of African American women, [is] shown in Table 1.2. The total range of scores for the exercise subscale for African American women is shown in Table 1.3 along with the range of scores for the two income groups.

▶ *(Note that upper ranges of the total HPLP scores were very similar for both groups, whereas lower-range scores differed; those below the poverty level had a large range of scores, along with greater variability. Exercise scores, on the other hand, showed identical lower-range scores but differed on the upper range. And, although not reported, the two means are significantly different.)*

Table 1.2 Range of Scores for Total HPLP

Group	*Range*	*M*	*SD*
Total women ($n = 58$)	1.50–3.56	2.71	.48
Above poverty ($n = 32$)	1.90–3.56	2.85	.40
Below poverty ($n = 26$)	1.50–3.53	2.51	.51

Note: HPLP = Health-Promoting Lifestyle Profile; *M* = mean; *SD* = standard deviation.

Table 1.3 Range of Scores for the Exercise Subscale

Group	*Range*	*M*	*SD*
Total women ($n = 58$)	1.00–3.60	2.05	.81
Above poverty ($n = 32$)	1.00–3.60	2.30	.77
Below poverty ($n = 26$)	1.00–3.40	1.72	.76

Note: *M* = mean; *SD* = standard deviation.

Discussion

Findings from this study indicate that African American women living above poverty level engage in more health-promoting behaviors than do African American women living below poverty level.

▶ *(Note that the conclusion is inappropriate. Health-promoting behaviors were not observed; they were reported. Because validity measures for the questionnaire were not offered, we don't know the extent to which test items accurately reflect behavior. Moreover, we don't know the accuracy of the self-reports.)*

The mean scores of African American women on the exercise subscale of this study were also low. . . . The five-item scale based on self-reported exercise behaviors was ranked as 1 (*never*), 2 (*sometimes*), 3 (*often*), and 4 (*always*). Thus, even if a woman never exercised, the mean score would be 1.0.

▶ *(Note that scores might have been higher if all forms of exercise, not just recreational, were reported.)*

Researchers for a previous study found that there was no significant difference in exercise behaviors between young, middle-aged, and older adults. In fact, exercise had the lowest score of all six variables on the HPLP. . . .

Implications for holistic nursing, primary prevention, and primary care practice include acknowledging and encouraging African American women who do use health-promoting behaviors, especially exercise or some form of physical activity in their daily life. . . .

For African American women living below poverty level, specific prescriptions for exercise need to be developed that are culturally specific, practical, and inexpensive. . . .

▶ *(This is misleading, because it implies that exercise is a main factor that accounts for the difference in HPLP between the two groups. First, other subscale differences were not considered, and they may have differed as well. Second, factors that are not associated with one group being at below poverty level and another group at above poverty level may account for the differences in health-promoting behaviors [e.g., health status, intelligence, childhood rearing, marital history, closeness with family members]. Third, we don't know the extent to which "encouraged" participation and help in completing the questionnaire unintentionally affected responses and did so for one group more than the other. Thus, if groups were matched on all non-poverty level variables and were tested by a naive [with respect to the purpose of the study] individual, it would be possible to reach a valid conclusion.)*

STUDY EXAMPLE 1.2: "BEHAVIORAL RESPONSES OF NEWBORNS OF INSULIN-DEPENDENT AND NONDIABETIC, HEALTHY MOTHERS"

This next study concerns developmental differences between two groups of newborn babies: those born to insulin-dependent diabetic mothers and those born to mothers without diabetes. Because measures were taken on two different days, the statistical analyses involved more than just *t* tests. However, those results can be summarized without getting into these statistics, which we'll cover in a later chapter.

The Study

Pressler, J. L., Hepworth, J. T., LaMontagne, L. L., Sevcik, R. H., & Hesselink, L. F. (1999). Behavioral responses of newborns of insulin-dependent and nondiabetic, healthy mothers. *Clinical Nursing Research*, *8*(2), 103–118. Copyright © 1999 by Sage.

Mother–child interactions begin before and shortly after birth. The mother's reactions in part depend on the physical condition of the newborn. Of some concern is the newborn baby of a mother with insulin-dependent diabetes. Several studies have shown such babies to be slower to develop, as measured by the Neonatal Behavioral Assessment Scale (NBAS). Other studies have found no differences between these babies and ones born to healthy mothers. These earlier studies focused on the mother's control of glucose during pregnancy. But there is a need

to consider other aspects of pregnancy and delivery as well. The present study attempted to control for "type of delivery, labor and delivery medications, parity, race and ethnicity, and maternal education" to evaluate development of newborns of insulin-dependent diabetic mothers (NDMs) and to healthy mothers.

Method

Participants

A convenience sample of 40 term newborns whose mothers had been cared for within the antenatal clinics of a metropolitan medical center and whose maternal diabetes had been closely monitored for glucose control and complications throughout pregnancy, was matched with 40 newborns of nondiabetic, healthy mothers (controls) for type of delivery, labor and delivery medications, parity, race and ethnicity, and maternal education. . . . Any pregnant diabetic mother who experienced complications related to diabetes or whose glucose levels were unable to be managed by her obstetrician was not considered eligible for this study.

▶ *(Note that nothing is said about how many clinics were involved nor whether both groups of participants were drawn from all or just one clinic. If prenatal care is more efficient in one clinic, then this might be a factor that further differentiates the groups.)*

Twenty-three of the NDMs and 18 of the controls were male, and 17 of the NDMs and 22 of the controls were female. Newborns' 5-minute Apgar scores were at least 7, admission physical examinations were assessed to be normal, no major congenital anomalies were evident, and none were placed under different from ordinary assessment protocols. Only newborns who were 37 to 42 weeks gestation . . . were studied, and there was no difference in gestational age between groups, $t(78) = .25$, $p = .80$. The NDMs had an average gestational age of 39.08 weeks ($SD = 1.47$) compared with controls' mean age of 39.15 weeks ($SD = 1.19$). Mean birth weights were 3,660 g ($SD = 542$) for the NDMs and 3,239 g ($SD = 298$) for the controls, $t(60.5) = 4.31$, $p < .01$.

▶ *(Note that the degrees of freedom [*df*] for the weight* t *ratio is 60.5, but it should be 40 + 40 − 2 = 78. When the* t *is calculated using the means and standard deviations [*SDs*] given and 40 in each group, the* t *is as presented as 4.31. Therefore, the reported* df *of 60.5 is a typographical error. It is equally important to note that not only do the two mean weights differ significantly but so does variability. This can be checked by forming a ratio between* $542^2/298^2 = 3.308$. *According to an* F *table [to be covered in a later chapter], the 3.308 is significant at* p $< .01$. *This indicates that there was a greater range of weight in the NDM group than in the healthy mothers group.)*

Of the 40 matched newborn pairs studied, 26 pairs were delivered vaginally and 14 were delivered by cesarean birth. Twelve mothers received either no medications before delivery or only a local anesthetic, 4 received analgesics and a local anesthetic, 19 received epidural anesthetics and analgesics, and 5 received general anesthesia and analgesics. Fifteen were firstborn and 25 were later-born pairs; 34 were Caucasian and 6 were African American.

▶ *(Note that these figures for type of delivery add up to 40. This indicates that both groups were perfectly matched on one critical potentially confounding factor—method of birth delivery.)*

The Neonatal Behavioral Assessment Scale

The Brazelton NBAS . . . was used to assess newborn behavior. At present, the NBAS is the more reliable measure of the neonate's interactive responses on a wide range of behavior in an interactional process. . . . The exam takes approximately 25 to 45 minutes to administer. In addition to the 20 reflex items, the NBAS exam contains 27 behavioral variables, scored for optimal performance, with 8 of the behavioral variables receiving a second score for modal performance, bringing the total number of items to 55.

. . . A modal response score can be created that represents the most frequently occurring orientation responses. . . . Higher scores on the reflex functioning dimension of the NBAS indicate poorer performance. Higher scores on the remaining behavioral dimensions and the modal response score indicate better performance. For the current study, the optimal score for the range of state dimension, which is an average score of those items contained within this dimension, was considered to be 3.

▶ *(Note that the test used to measure the infants is known to be reliable and valid.)*

Procedure and Data Analysis

The three NBAS examiners involved in this study received formal certification for NBAS administration and scoring. Interscorer reliability among these certified examiners was achieved when they were in at least 90% agreement on three tests . . . . Prior to data collection, interobserver agreement, allowing for a one-point discrepancy, was .93. All NBAS examiners maintained greater than a .90 level of reliability when intermittent reliability checks were completed during data collection.

All insulin-dependent diabetic mothers whose newborns met the selection criteria were approached for participation during early labor or during the first 12 hours postdelivery. After obtaining each NDM participant, control newborns were consecutively obtained when matched for the five maternal variables. All mothers gave informed consent for their newborns' participation in the study. All newborns

were examined using the NBAS on the first and second days, at 12 to 24 hours and again at 36 to 48 hours of postnatal life. . . .

▶ *(Note that all examiners were well qualified and that data were included only when there was a high degree of interscorer agreement. This leads to reliable data. Note, too, that the healthy mothers had to be recruited after the diabetic mothers were recruited to ensure adequate matching. It would have been useful, however, to know how long after the diabetic mothers were recruited the healthy mothers were recruited, to estimate the time lapse between testing the NDM babies and the babies of healthy mothers.)*

Results

There were no differences between the groups on socioeconomic status (SES) as evaluated using Hollingshead's . . . four factor index of social status $t(78) = .98$, $p = .33$ . . . or on the first month seen by a physician during the pregnancy, $t(78) = 1.07$, $p = .29$. Diabetic mothers had an average SES score of 33.4 ($SD = 16.4$) and were first seen in the clinics at month 3.05 ($SD = 1.60$). The control mothers had an average SES score of 30.0 ($SD = 14.0$) and were first seen at month 3.42 ($SD = 1.53$).

As a reflection of the close monitoring of the diabetic mothers, differences were found between the groups on the number of prenatal visits, $t(61.6) = 3.92$, $p < .01$; the number of hospitalizations, $t(39.0) = 8.20$, $p < .01$; and the average number of prenatal visits, $t(62.4) = 3.85$, $p < .01$. The diabetic mothers had more prenatal visits ($M = 13.12$, $SD = 4.99$), more hospitalizations ($M = 1.52$, $SD = 1.18$), and more prenatal visits per month of prenatal care ($M = 1.92$, $SD = .64$) than control mothers ($M = 9.58$, $SD = 2.82$; $M = 0.00$, $SD = 0.00$; and $M = 1.47$, $SD = 37$, respectively).

▶ *(Note that the* df*s reported for all three* t *tests are wrong and reflect typographical errors. Each should have 78* df*s, and when these t values were recalculated with the correct* df*, the same* t*'s were obtained. Note, too, that the additional visits might put the diabetic group at an advantage. If the mothers-to-be were too sick, on the other hand, they would not have carried to term and would not have given birth to babies who weighed more than those born to healthy mothers.)*

The mean [*M*] scores and *SD*s are presented in Table 1.4. . . . Only the significant results are presented here.

Significant differences were found between NDMs and their matched controls on motor processes, $t(78) = 2.951$, $p < .01$, and reflex functioning—$t(78) = 5.551$, $p < .01$.

Table 1.4 Mean Scores and Standard Deviations by Maternal Health Status and Time of Testing

		NDMs		*Controls*	
		Day 1	*Day 2*	*Day 1*	*Day 2*
Response decrement	*M*	6.58	6.38	7.20	5.85
	SD	1.43	.92	.97	.53
	n	6	6	6	6
Orientation	*M*	7.18	7.27	7.35	7.53
	SD	.83	.86	1.02	.83
	n	37	37	37	37
Range of state	*M*	1.95	1.76	1.93	1.93
	SD	.80	.91	.60	.90
	n	40	40	40	40
Motor processes	*M*	4.96	5.38	5.31	5.73
	SD	.76	.70	.73	.51
	n	40	40	40	40
Autonomic stability	*M*	6.07	6.37	6.18	6.41
	SD	1.07	1.00	.93	.77
	n	40	40	40	40
Regulation of state	*M*	3.98	4.16	4.78	4.51
	SD	2.12	2.04	2.19	2.04
	n	40	40	40	40
Reflex functioning	*M*	1.53	.93	.73	.13
	SD	1.26	1.54	1.11	.52
	n	40	40	40	40
Modal responses	*M*	6.57	6.89	6.84	7.20
	SD	1.13	.92	1.18	.91
	n	35	35	35	35

Note: M = mean; *n* = sample size; *NDM* = newborn of insulin-dependent diabetic mother; *SD* = standard deviation.

▶ *(Actually,* F *values were presented, but* $F = t^2$*, so the values presented are the square root of F. Also note that, because of sizes of the standard deviations, not all babies in the NDM group were slower. Some fit within the normal range.)*

The NDMs' motor performances were poorer than comparison newborns' motor performances, and NDMs also had significantly more abnormal reflexes than their matched controls. . . .

Time of testing effects were found for response decrement, . . . autonomic stability, . . . reflex functioning, . . . and modal performance. . . . Except for the response decrement dimension, there was better performance on Day 2 than on Day 1 for motor processes, autonomic stability, reflex functioning, and modal performance. . . . None of the effects for orientation, range of state, or regulation of state were significant.

Discussion

Controlling for the maternal variables identified in this study, term NDMs evidenced a number of behaviors in motor processes and reflex functioning that would portray them as more lethargic and listless than their matched healthy controls. Also, the means for the NDMs on all behavioral dimensions were in the direction that one would expect for a tired and/or listless baby. The results also revealed that motor processes, autonomic stability, reflex functioning, and modal performance improved for both NDMs and control newborns after 24 hours, but the performance of NDMs on these dimensions was not as robust as that of healthy controls.

The poorer motor and reflex functioning of the NDMs might pose a challenge for parents. Mothers of NDMs need to be informed that their newborns might exhibit different behaviors from typical, healthy newborns and, consequently, present natural behavioral barriers toward effective parent–infant interaction. . . .

A limitation of the current study was that actual mother–infant interaction was not measured. Future investigations need to be conducted using both behavioral and interactive measures so that a knowledge base can be developed, with respect to any differences that might exist in newborn behavioral and early postnatal interactions of NDMs and their mothers compared with newborns of nondiabetic, healthy mothers. . . .

Finally, the findings of this study support the research of other investigators who found that even closely monitored and/or well-controlled diabetes during pregnancy can adversely affect NDMs' early behaviors. . . . It also emphasizes the importance of considering a number of maternal background variables in the design of the future studies, when the aim is to understand the influence of maternal diabetes on newborns' behavioral responses.

▶ *(The basic conclusion of the study is that newborns of diabetic mothers can be slower in certain reflex and motor behaviors soon after birth. The implication is that the mother's physiology, particularly blood sugar level, has an adverse effect on the prenatal baby. Diabetic women not only take insulin but also carefully monitor how much they eat from each food group. This is part of the treatment of diabetes. There is, however, a psychological side as well, a possible increase in stress as one tries to cope. To the extent that there is increased stress,*

there is an increase in stress hormones, which also might affect the fetus. What is important is that the effect on newborns to some part of having diabetes might be responses to slower development during the first 3 to 4 days that were the focus of the study. The samples were very well matched, ruling out as possible confounds a host of variables that might be associated with prenatal care and delivery. On these bases, the conclusion is justified. The only question is whether both groups of mothers were drawn from the same clinics. To the extent that they were, the conclusion holds up. To the extent that they were not, there might be differences in quality of prenatal care that might contribute to the observed differences in babies and that might even be unrelated to diabetes.)

OVERVIEW OF THE REMAINING CHAPTERS

For the remainder of this book, each chapter describes a research design and presents representative articles. A summary of the designs and their description is provided in Box 1.3. Although earlier chapters describe research designs that tend to be simpler than later chapters, each chapter is self-contained. Feel free to jump around, read the chapters out of order, or focus just on the designs that are of most interest to you. Our goal is to help you understand research based on a variety of designs and enable you to read research articles in a more critical and personally meaningful way.

Box 1.3 Study Designs by Chapter

Ch.	*Design*	*Description*
2	Case Studies	Qualitative research based on an in-depth study of an individual or small group
3	Narrative Analysis	Qualitative analysis of a chronologically told story, with emphasis on how elements are sequenced and evaluated
4	Surveys	Research that collects descriptive information about the members of a population in a standardized fashion
5	Correlation Studies	Observational research that describes relationships among quantitative variables
6	Regression Analysis Studies	Research that examines the prediction of a quantitative outcome from one or more predictor variables

(Continued)

Box 1.3 (Continued)

Ch.	*Design*	*Description*
7	Factor-Analytic Studies	Research that develops or confirms theories about unobserved hypothetical variables by examining correlations among observed or measured variables
8	Discriminant Analysis Studies	Research that examines the prediction of a categorical outcome from one or more predictor variables
9	Two-Condition Experimental Studies	Design in which subjects are randomly assigned to one of two experimental conditions, and the impact of the conditions on a dependent variable is assessed
10	Single Classification Studies	Study that evaluates the impact of differing conditions, defined by the levels of a single categorical independent variable, on a dependent variable
11	Factorial Studies	Study that evaluates the impact of differing conditions, defined by combinations of two or more independent variables, on a dependent variable
12	Quasi-Experimental Studies	Study in which participants are not randomly assigned to levels of the independent variable
13	Longitudinal Studies	Study in which individuals or comparable groups of individuals are assessed over time

READING RESEARCH ARTICLES (BUT NOT FROM START TO FINISH)

Before moving on to studying specific designs, let's take a moment to discuss the process of reading a research article. Most readers approach an article in a linear fashion, reading from beginning to end. Articles can be quite dense, and there is nothing more frustrating than getting halfway through an article, only to realize that it is not relevant for your work. As an alternative method for quickly scanning an article, consider the following approach:

1. Look at the name of the journal. Rigorous and selective journals will typically publish research that has undergone meticulous peer review.
2. Read the article title. This will give you general information about the article.
3. Read the abstract. This will give you a quick and succinct overview of the article.
4. Read the first paragraphs of the introduction. This will generally describe why the subject matter is important.
5. Read the last paragraphs of the introduction. This will generally describe what the author plans to do and why.

6. Read the first paragraphs of the discussion. This will often include a short summary of what was found in the results.
7. Read the last paragraphs of the discussion. This ties the work together and puts it in context.
8. If the article is still interesting to you, read it carefully from beginning to end. Use the methods discussed in this book to evaluate the quality of the research and appropriateness of the conclusions.

Although authors vary in their writing—and this approach will not work every time—it can often save you time and help you grasp the overall structure and purpose of the article. Note, however, that the articles in this book have *already* been abstracted, so the method described here should only be used when approaching complete articles in journals.

BIBLIOGRAPHY

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Manual for the Beck Depression Inventory-II*. San Antonio, TX: Psychological Corporation.

SUGGESTED READINGS

Creswell, J. W. (2009). *Research design: Qualitative, quantitative, and mixed methods approaches* (3rd ed.). Thousand Oaks, CA: Sage.

Graziano, A. M., & Raulin, M. L. (2006). *Research methods: A process of inquiry* (6th ed.). Boston: Allyn & Bacon.

Isaac, S., & Michael, W. B. (1995). *Handbook in research and evaluation* (3rd ed.). San Diego, CA: EdiTS.

Kazdin, A. E. (2002). *Research design in clinical psychology* (4th ed.). Boston: Allyn & Bacon.

Mason, E. G., & Bramble, W. J. (1997). *Handbook in research and evaluation* (3rd ed.). Dubuque, IA: Brown & Benchmark.

McMillan, J. H., & Schumacher, S. (2000). *Research in education: A conceptual introduction* (5th ed.). White Plains, NY: Longman.

Nunnaly, J., & Bernstein, I. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Lawrence Erlbaum.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2001). *Quasi-experimentation: Design & analysis issues for generalized causal inference*. Boston: Houghton Mifflin.

Shultz, K., & Whitney, D. (2004). *Measurement theory in action: Case studies and exercises*. Thousand Oaks, CA: Sage.

Sproull, N. L. (2003). *Handbook of research methods: A guide for practitioners and students in the social sciences* (2nd ed.). Lanham, MD: Scarecrow Press.

Vockell, E. L., & Asher, J. W. (1995). *Educational research* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.