INSTRUCTOR'S MANUAL
WITH TESTS

to accompany

Beginning Behavioral Research: A Conceptual Primer

(Seventh Edition)

Ralph L. Rosnow
Robert Rosenthal

Michael E. Greenberg, Ph.D.
Shippensburg University
and
Beth A. Greenberg, MA, MPA
Harrisburg Area Community College
and Shippensburg University

Prentice Hall
is an imprint of

PEARSON

10 9 8 7 6 5 4 3 2 1

ISBN 10: 0-205-87189-5
ISBN 13: 978-0-205-87189-6

Printed in the United States of America

CONTENTS

PART III DESIGN AND IMPLEMENTATION

PART IV DESCRIBING DATA AND MAKING INFERENCES

PART V STATISTICAL TESTS

PREFACE

The purpose of this Instructor's Manual with Tests is to introduce undergraduate students to methods used by behavioral scientists, as well as to reacquaint new and experienced instructors to them. Each chapter of the text has an outlined summary, and most are followed by classroom-tested instructional activities, discussion topics, and demonstration exercises. Finally, there are multiple-choice and short essay questions that cover the core material for each chapter. The multiple-choice questions include the page numbers in the text where the answers are found.

Update to the Seventh Edition (7e)

Building on the strong foundation of the existing Instructor's Manual by David B. Strohmetz, Monmouth University, and Eric K. Foster, Temple University, we sought to align the changes in topics and chapters with the content. The Seventh Edition poses questions at each section heading and we updated the Instructor's Manual accordingly. New questions have been added, again, generally following the changes and additions to this edition.

Michael Greenberg & Beth Greenberg – Shippensburg University, May 2012.

CHAPTER 1: *BEHAVIORAL RESEARCH AND THE SCIENTIFIC METHOD*

CHAPTER OUTLINE

I. Why Study Research Methods and Data Analysis?
- A. The term "researching" (i.e., exploring a problem systematically) is traditionally called the scientific method in college science courses.
  1. This "method" is used in all scientific fields.
  2. However, its applications vary from one discipline to another.
- B. Why should we know the scientific method or study techniques of research?
  1. We can enhance our understanding of the influence that science has on our lives.
  2. We can learn to differentiate between good science and pseudoscience.
  3. We can acquire information and skills useful in our daily lives.
  4. We can learn about the limits of particular studies and methods.
  5. We may find that studying and doing research can be an exciting career.

II. What Alternatives Are There to the Scientific Method?
- A. Charles Sanders Peirce (1839–1914) described four distinct strategies for formulating strongly held beliefs.
- B. The four strategies for the "fixation of belief."
  1. **Method of tenacity** is clinging stubbornly and mindlessly to claims or beliefs just because they have been around a while.
  2. **Method of authority** is the acceptance of an idea as being valid because someone in a position of power or authority states it.
  3. The **a priori method** is the use of one's individual powers of reason and logic to make sense of the world.
  4. The **scientific method** provides a framework with which to draw on independent realities to evaluate claims.

III. How Do Scientists Use Empirical Reasoning and the Scientific Method?
- A. The scientific method involves the use of **empirical reasoning**.
- B. Empirical reasoning is a combination of logic, carefully organized observation, and measurement.
- C. It is the use of empirical reasoning that all scientists have in common, despite differences in the particular methods of empirical inquiry they may employ.
- D. Empirical reasoning entered into behavioral science during the late nineteenth century when individuals such as Wilhelm Wundt (1832–1920) and William James (1843–1910) began employing the scientific method utilized by physicists and biologists to study psychological behavior.
- E. Francis Galton (1822–1911) demonstrated the application of empirical reasoning to questions thought to lie completely outside of science.

IV. Applications in Behavioral Research

A. Empirical reasoning has been applied to questions about human nature, cognition, perception, and behavior.

B. Stephen J. Ceci and his colleagues employed empirical reasoning to investigate the accuracy of children's eyewitness testimony.

C. Solomon Asch used empirical reasoning to study conformity and the reasons why people go along with certain consensual opinions.

V. How Do Extraempirical Factors Come into Play?

A. Although the scientific method is distinguished by its reliance on the primary use of empirical procedures, extraempirical factors also play an important role in ascertaining what is true.

B. Aesthetic considerations play a part.

.

C. Opinions and arguments are articulated in the accepted **rhetoric** (**rhetoric of justification)** of the particular field they represent.

1. Rhetoric includes specialized terms and structure of reporting.
2. Peer-reviewed journals rely upon this rhetoric.

D. Researchers have a penchant for poignant analogies and metaphors for visualizing one thing in terms of another (i.e., **perceptibility**).

VI. What Does Behavioral Research Cover?

A. **Behavioral Research** is an umbrella term that includes covers the use of empirical reasoning (viz., careful logic, organized observation, and measurement) from different methodological vantage points in an effort to understand how and why people act, perceive, feel, and think as they do in a variety of disciplines such as psychologists, behavioral economists, political scientists, sociologists, and cultural anthropologists.

B. The objective of behavioral and social science is to describe and explain how and why humans think, feel, and behave as they do.

C. To develop a more complete and integrated picture of human nature, behavioral and social scientists have come to embrace **methodological pluralism,** which means that by necessity, researchers use different tools and designs (different methods) because each is limited in some way, yet each method represents and reflects a particular perspective on the phenomenon of interest and the multifaceted complexity of human nature.

VII. How Does Research Go From Descriptive to Relational to Experimental?

A. Descriptive conclusions tell us *how things are*.

1. The goal of **descriptive research** is the careful mapping out of a situation or set of events.
2. Causal explanations are not of direct concern except perhaps speculatively.
3. This orientation is often considered a necessary first step in the development of a program of research because it establishes the logical and empirical foundation of any future undertaking.

4. Descriptive research is rarely regarded as sufficient as it does not allow one to address questions concerning why something happens or how what happens is related to other events.

B. Relational (or "correlational") conclusions tell us *how things are in relation to other things*.
1. Relational (or correlational) research involves measuring and relating two or more variables or conditions.
2. Based on coordinated observations, one should be able to make a quantitative statement concerning the relationship, or correlation, between the variable of interest.
a. Are *X* and *Y* significantly related?
b. What is the pattern of the relationship (e.g., linear or nonlinear)?
c. What is the strength of the relationship?

C. Experimental conclusions tell us *how things are and how they got to be that way*.
1. The objective is the identification of causes (i.e., what leads to what) through the manipulation of conditions thought to be responsible for the effect.
2. Relational research rarely provides causal explanations, and then only under very special circumstances.

VIII. What are the Characteristics of Good Researchers?
A. Enthusiasm
B. Open-mindedness
C. Common sense
D. Role-taking ability
E. Creativity and inventiveness
F. Confidence in one's own judgment
G. Ability to communicate
H. Care about details
I. Integrity and honest scholarship

LECTURE IDEAS AND ACTIVITIES

1. To demonstrate the pervasiveness of science in modern society as well as the utility of understanding the process of science, assemble a collection of articles that report on recent scientific findings. The science section of the Tuesday *New York Times* is particularly useful for finding such articles (http://www.nytimes.com). Another Internet resource is *Science Daily* (http://www.sciencedaily.com). Discuss the findings reported in these articles, emphasizing how an understanding of the scientific process can help one better evaluate or question the findings or conclusions reported in the media.

2. Jacobson, Mulick, and Schwartz (1995) discuss how the reliance on pseudoscientific findings has led to the acceptance by professionals of some therapeutic treatments that appear to have negligible, if any, benefit for the afflicted individual. Jacobson et al. argue that one example of the reliance on pseudoscientific research practices to establish the efficacy of a therapeutic

intervention is the controversial case of facilitated communication. Jacobson et al. describe the disparity between the controlled, scientific research studies that have found very little, if any, support for this type of intervention with autistic individuals and its unquestioned acceptance by its proponents. Jacobson et al. discuss possible reasons why proponents of facilitated communication have rejected sound scientific practices in favor of practices that can be described as representing pseudoscience. Not surprisingly, this article sparked debate concerning whether scientific practices can really establish the efficacy of facilitated communication (e.g., Allen & Allen, 1996; Biklen, 1996; Fernald, 1996; Jacobson et al., 1996; Knox, 1996). You might want to assign these articles and have your students debate the criteria that one should use to establish the effectiveness of a treatment intervention. You may also want to discuss whether treatments that have become popular based solely on pseudoscientific evidence are really that detrimental to society as a whole or to the individuals they are intended to help. In other words, is it always necessary to establish the efficacy of a treatment intervention using practices that can be characterized as "good science"?

Allen, B., & Allen, S. (1996). Can the scientific method be applied to human interaction? *American Psychologist, 51,* 986.

Biklen, D. (1996). Learning from the experiences of people with disabilities. *American Psychologist, 51,* 985–986.

Fernald, D. (1996). Tapping too softly. *American Psychologist, 51,* 988.

Jacobson, J. W., Mulick, J. A., & Schwartz, A. A. (1995). A history of facilitated Communication: Science, pseudoscience, and antiscience. *American Psychologist, 50,* 750-765.

Jacobson, J. W., Mulick, J. A., & Schwartz, A. A. (1996). If a tree falls in the woods... *American Psychologist, 51,* 988–989.

Knox, L. A. (1996). The facilitated communication witch-hunt. *American Psychologist, 51,* 986–987.

3. More information on the life of Charles Sanders Peirce as well as hypertext versions of his writings are available at a website dedicated to this American philosopher (http://www.peirce.org).

4. Before discussing Peirce's methods of "fixing belief," have students write down five things they believe to be true. Once they have completed their lists, have each student share his or her list with another student. As one student reads each "truth" from his or her list, the student's partner should simply ask, "Why do you believe that this is true?" to each item, recording the student's response. As a class, discuss the nature of the arguments that were used to fend off the challenges to the veracity of the student's beliefs. This exercise easily leads into a discussion of Peirce's methods of "fixing belief." You may want to categorize the types of arguments the students used to justify the veracity of their beliefs using Peirce's four methods of fixing belief.

5. To help students to critically consider the underlying foundation of claims of veracity, you might incorporate the following writing assignment into discussion of Peirce's methods of

"fixing belief." The assignment involves the analysis of the famous 1897 editorial entitled, "Yes, Virginia, There Is a Santa Claus." This editorial is actually a reply to a letter from an eight-year-old girl named Virginia O'Hanlon who was asking the editor of the now defunct newspaper, the *New York Sun*, the truth about the existence of Santa Claus. The editor, Francis Church, utilizes three of Peirce's methods to establish that Santa Claus does exist while at the same time rejecting the scientific method as a means for answering this question. Have students write a two- to three-page paper evaluating this exchange between Virginia and Church in terms of Peirce's four methods.

Virginia's letter and Church's subsequent reply is widely available in the public domain on the Internet. Here is the exchange between the two individuals:

> Dear Editor: I am 8 years old. Some of my little friends say there is no Santa Claus. Papa says, 'If you see it in *The Sun* it's so.' Please tell me the truth, is there a Santa Claus?"
>
> Virginia O'Hanlon
> 115 WEST 95TH ST.
>
> Virginia, your little friends are wrong. They have been affected by the skepticism of a skeptical age. They do not believe except they see. They think that nothing can be which is not comprehensible by their little minds. All minds, Virginia, whether they be men's or children's, are little. In this great universe of ours man is a mere insect, an ant, in his intellect, as compared with the boundless world about him, as measured by the intelligence capable of grasping the whole of truth and knowledge.
>
> Yes, Virginia, there is a Santa Claus. He exists as certainly as love and generosity and devotion exist, and you know that they abound and give to your life its highest beauty and joy. Alas! how dreary would be the world if there were no Santa Claus! It would be as dreary as if there were no Virginias. There would be no childlike faith then, no poetry, no romance to make tolerable this existence. We should have not enjoyment, except in sense and sight, The eternal light with which childhood fills the world would lie extinguished.
>
> Not believe in Santa Claus! You might as well not believe in fairies! You might get your papa to hire men to watch in all the chimneys on Christmas Eve to catch Santa Claus, but even if they did not see Santa Claus coming down, what would that prove? Nobody sees Santa Claus, but that is no sign that there is no Santa Claus. The most real things in the world are those that neither children nor men can see. Did you ever see fairies dancing on the lawn? Of course not, but that's no proof that they are not there. Nobody can conceive or imagine all the wonders there are unseen and unseeable in the world.
>
> You tear apart the baby's rattle and see what makes the noise inside, but there is a veil covering the unseen world which not the strongest man, nor even the united strength of all the strongest men that ever lived, could tear apart. Only faith, fancy, poetry, love, romance can push aside that curtain and view and picture the supernal beauty and glory beyond. Is it all real? Ah, Virginia, in all this world there is nothing else real and abiding. No Santa Claus! Thank God! he lives, and he lives forever. A thousand years from now,

Virginia, nay, ten times ten thousand years from now, he will continue to make glad the heart of childhood.

Strohmetz, D. B. (2003, March). *Ways of Knowing There Is A Santa Claus*. Poster session presented at the 17th annual conference on Undergraduate Teaching of Psychology: Ideas and Innovations, March, 2003, Ellenville, NY.

6. Keith Stanovich (2001) tells the story of how Francesco Szizi, an astronomer, tried to refute Galileo's claim that there were moons orbiting Jupiter. Rather than looking through Galileo's telescope, Szizi rejected the possibility of Galileo's observation being true through the use of reasoning based on "common sense" (p. 9). Students will be amused by the absurdity of Szizi's argument as being considered a perfectly acceptable alternative to the use of systematic observation in the formation of beliefs and explanations concerning the world. However, point out to them that many of their own beliefs about the world may also be the product of reasoning similar to Szizi's. To demonstrate this point, have students complete Vaughn's (1977) Test of Common Beliefs. Students may find that many of the beliefs they have about human behavior have been shown by scientists to be inaccurate. Discuss how students may have originally formulated these misconceptions about behavior and what the possible implications would be to society if these types of beliefs were never challenged.

Stanovich, K. E. (2001). *How to think straight about psychology (6th ed.).* Boston, MA: Allyn and Bacon.

Vaughn, E. D. (1977). Misconceptions about psychology among introductory psychology students. *Teaching of Psychology, 4,* 138–141.

7. Art Kohn (1999) employs a variation of the "Monty Hall" problem to demonstrate the superiority of empirical reasoning over intuition (i.e., the method of tenacity). Kohn places a $1 bill in one of three envelopes, seals all three envelopes, and then shuffles the envelopes such that one no longer knows which envelope contains the dollar bill. Kohn asks for a volunteer to select one of the envelopes, stating that the person can keep the dollar if it is in the envelope selected. Kohn privately opens the other two envelopes, showing the class an envelope that does not contain the dollar bill. Before the volunteer opens his or her envelope, Kohn offers to switch envelopes with the volunteer. The question posed to the class is, should the volunteer switch the envelopes? Most will say that the volunteer should stay with her original selection, but will only be able to explain their reasoning through a "gut sense" or intuition rather than through empirical reasoning.

Kohn then has the class empirically evaluate their intuitive belief by conducting an experiment. Students form pairs in which one student is the experimenter and the other the research subject. Each experimenter is given a data sheet with four columns and twenty rows. The four columns are labeled "Correct Answer," "Subject's Choice," "Stay/Switch," and "Win/Lose." The experimenter fills in the rows in the "Correct Answer" column with a random assortment of the letters "A," "B," and "C." The experimenter is instructed to ask the subject to guess "A," "B," or "C." The subject is told which letter was not the correct one and then given

the opportunity to either stay with the original guess or switch to the nonchosen letter. The experimenter records the subject's decision as well as whether or not the subject ultimately made the right choice. After all the pairs of students have completed the experiment, Kohn compiles the class's results, determining the proportion of the time that the subject won (i.e., correctly selected the letter) by staying with the original selection as compared to the proportion of the time the subject won by switching.

Kohn has found that most students prefer the "staying strategy" but when looking at the overall results, they were more likely to win if they chose the "switching strategy." Kohn gives the mathematical rationale for why switching is better (because the odds are 2/3 that the experimenter chose the envelope containing the dollar and that these odds do not change even when the experimenter reveals an empty envelope).

This activity is useful for not only demonstrating the advantage of the scientific method over the method of tenacity in knowing our world, but also how truly "tenacious" the method of tenacity can really be. After having the class empirically demonstrate the superiority of the switching strategy over the staying strategy, offer the original volunteer again the opportunity to either keep his or her unopened envelope or switch envelopes. Despite the class's findings as well as the explanation for why switching is the best decision, volunteers frequently wish to stay with their original choice based on their "gut instinct." Needless to say, in these situations you will usually not lose a dollar to the student. Point out the potential cost that choosing to rely solely on intuition rather than empirical reasoning can have for the individuals (in this case, the cost is a dollar).

Kohn, A. (1999). Defying intuition: Demonstrating the importance of the empirical technique. In M. E. Ware and C. L. Brewer (Eds.), *Handbook for teaching statistics and research methods (2nd ed.)* (pp. 179–181). Mahwah, NJ: Lawrence Erlbaum.

8. The importance of rhetoric in science is illustrated in an article by Robert Madigan, Susan Johnson, and Patricia Linton (1995). These authors argue that "APA style" is not merely a set of writing rules. Rather, it reflects the rhetoric or language by which psychologists formulate their "ways of knowing." Madigan et al. assert that it is important for newcomers to the discipline of psychology to learn to communicate using this style as it reflects the accepted discourse as well as the history of psychology as a science. For example, among the important characteristics inherent in the "APA style" is the use of a story schema to describe one's use of systematic observations to address research questions, the use of hedged wording when drawing conclusions from these observations, and the emphasis on the empirical process rather than the individual through the passive voice and depersonalized styles of disagreement. While the implications of their arguments have been criticized (see Brand, 1996; Josselson & Lieblich, 1996; Vipond, 1996), Madigan et al. (1996) argue that one must learn the rhetoric of psychology (via the "APA style") to have the ability to make an impact upon or contribution to the field of psychology.

Brand, J. L. (1996). Can we decide between logical positivism and social construction views of reality? *American Psychologist, 51,* 652–653.

Josselson, R., & Lieblich, A. (1996). Fettering the mind in the nature of "science." *American Psychologist, 51,* 651–652.
Madigan, R., Johnson, S., & Linton, P. (1995). The language of psychology: APA style as epistemology. *American Psychologist, 50,* 429–436.
Madigan, R., Johnson, S., & Linton, P. (1996). APA style: Quo vadis? *American Psychologist, 51,* 653–655.
Vipond, D. (1996). Problems with the monolithic APA style. *American Psychologist, 51,* 653.

9. To encourage students to think carefully about the importance of nine traits Rosnow and Rosenthal note as characteristic of good researchers, have your students discuss in groups which three traits they believe to be most essential in order to be a good researcher and why. Interesting discussions should ensue as each group will most likely have selected different traits and for different reasons. You may also want to discuss how one can acquire or develop these traits.

MULTIPLE-CHOICE QUESTIONS

1. Which of Peirce's methods refers to the use of common sense or the obvious to justify claims of belief?
   * a. the method of tenacity
   b. the method of authority
   c. the a priori method
   d. the scientific method (3)

2. According to Peirce, the most primitive strategy for forming a belief is the
   a. a priori method.
   b. method of authority.
   * c. method of tenacity.
   d. scientific method. (3)

3. John refuses to alter his opinion because it seems so obvious it must be correct. According to Peirce, John is basing his belief on the
   a. a priori method.
   b. method of authority.
   * c. method of tenacity.
   d. scientific method. (3)

4. When an explanation is accepted as true because it is advocated by an expert, this belief is based on the
   a. method of tenacity.
   * b. method of authority.
   c. a priori method.
   d. scientific method. (4)

5. While having the oil changed in his car, Zach accepts his mechanic's recommendation to have his fuel-injection system cleaned. Which of Peirce's methods is Zach relying on to know what maintenance his car requires?
   a. the scientific method

b. the method of tenacity
c. the a priori method
* d. the method of authority (4)

6. Which method of fixing belief is based on the use of pure reason and logic?
a. the method of authority
* b. the a priori method
c. the scientific method
d. the method of tenacity (5)

7. Susan rejects the claim of a salesperson about the benefits of a particular product because it simply does not make logical sense to her. Which of Peirce's methods is Susan using to evaluate the salesperson's claim?
a. the method of authority
* b. the a priori method
c. the method of tenacity
d. the scientific method (5)

8. The inherent limitation of pure reason is a problem associated with which of Peirce's methods of fixing belief?
a. the scientific method
b. the method of tenacity
c. the method of authority
* d. the a priori method (5)

9. The use of empirical reasoning is essential to which of Peirce's methods of fixing belief?
a. the method of authority
b. the a priori method
c. the method of tenacity
* d. the scientific method (5)

10. By systematically testing the different electrical systems in his car, Joshua is using which of Peirce's methods to confirm his belief that he may need a new car battery?
a. the a priori method
* b. the scientific method
c. the method of authority
d. the method of tenacity (5)

11. Kristen exclaims, "I want to see the evidence for myself before I'll accept your explanation!" Kristen's claim reflects the basic idea behind the
* a. scientific method.
b. method of authority.
c. a priori method.
d. method of tenacity. (5)

12. Which of the following represents how Peirce ordered his strategies for formulating beliefs, from most flawed to least flawed?
* a. method of tenacity, method of authority, a priori method, scientific method
b. method of authority, method of tenacity, a priori method, scientific method
c. a priori method, method of tenacity, method of authority, scientific method
d. method of tenacity, a priori method, scientific method, method of authority (3-5)

13. The use of observation and experience in inquiry characterizes
a. armchair reasoning.
* b. empirical reasoning.
c. a priori reasoning.
d. authoritative reasoning. (5-6)

14. Which of the following terms refers to the use of techniques based on observation or experience?
a. a priori
b. pseudoscience
* c. empirical
d. analogical rhetoric (5-6)

15. Which of the following refers to visualizing one thing in terms of another?
a. theoretical ecumenism
b. ad hoc hypothesis
* c. perceptibility
d. methodological pluralism (11)

16. Technical terms such as "hypotheses," "participant observation study," and "intercoder reliability" are examples of
a. methodology pluralism.
b. theoretical ecumenism.
* c. the rhetoric of justification.
d. pseudoscience. (10)

17. Kim is interested in what physical qualities people most desire in their prospective partners. She begins by interviewing married individuals, asking them what physical qualities most attracted them to their current partner. She then visits a bar and observes the physical qualities of those who leave with each other. Finally, she examines the personal ads in a local newspaper, noting the physical qualities that are most often mentioned in these ads. In her investigation of interpersonal attraction, Kim is employing
a. experimental methodology.
* b. methodological pluralism.
c. theoretical ecumenism.
d. an interdisciplinary approach. (8-9)

18. To develop a richer, more complete understanding of human behavior, researchers embrace _____ because they recognize that there is often more than one "right way" to view the causes of behavior.
* a. methodological pluralism
b. theoretical ecumenism

c. analogical thinking
d. empirical reasoning (8)

19. The objective of descriptive research is to determine
* a. what's happening.
b. what's related.
c. what caused it.
d. what does it affect. (12)

20. Interested in how teenagers interact when unsupervised, Cheryl decides to spend several Saturdays observing adolescents at a local mall. Cheryl's work can be BEST described as
a. experimental research.
b. relational research.
c. quasi-research.
* d. descriptive research. (12-13)

21. Amala spends several hours at a local playground observing how often the children engage in cooperative as well as competitive activities. Amala's investigation can best be described as
* a. descriptive research.
b. quasi-experimental research.
c. experimental research.
d. relational research. (12-13)

22. Dave is interested in whether one's support for prayer in schools is associated with one's religiosity. This type of question is most characteristic of
a. descriptive research.
b. laboratory research.
* c. relational research.
d. experimental research. (13)

23. Interested in the effect of outside employment on academic performance, a professor asks his students how many hours a week they work and compares this to current grade point averages. The professor's inquiry is an example of which broad research approach?
a. experimental
b. pseudoscientific
* c. relational
d. descriptive (13)

24. Which of the following questions is beyond the scope of relational research?
a. Are *X* and *Y* significantly related?
b. What is the form of the relationship between *X* and *Y*?
* c. Will changes in *X* cause changes in *Y*?
d. How strong is the relationship between *X* and *Y*? (13)

25. Mary suspects that a new violent afternoon TV show is the reason for her son's sudden increase in aggressive behavior towards his sister. Which research approach would best help Mary evaluate this suspicion?
a. correlational

b. descriptive
c. relational
* d. experimental (13)

26. To evaluate questions of causality, scientists must conduct
a. relational research.
* b. experimental research.
c. laboratory research.
d. descriptive research. (13-14)

27. Brian suspects that his new late-night cappuccino habit is the cause of his recent insomnia problems. Which research approach would best help Brian evaluate his suspicion?
a. anecdotal
b. relational
c. descriptive
* d. experimental (13-14)

28. Which of the following is NOT one of the orienting attitudes of scientists described in the text?
a. open-mindedness
b. confidence in one's own judgment
c. ability to communicate
* d. ability to be correct (15-16)

29. When they finished describing the events, John and Mary felt sure they knew how the events were related. What strategy should they use next?
a. Write their conclusions, in APA format, stating their observations and outlining the relationships between the events and the causes of the events.
* b. Conduct relational research to determine the relationships between the conditions or variables.
c. Use the a priori method.
d. Develop more open-mindedness. (13)

SHORT ESSAY QUESTIONS

1. Discuss three reasons of the five reasons mentioned in the textbook why it is beneficial for one to learn and know about the scientific method.

2. What are Peirce's four strategies for formulating explanations? Which of these strategies is the least desirable? Why? Which is the most desirable? Why?

3. Why is an accepted rhetoric one of the features of the scientific method? Describe some aspects of this rhetoric.

4. What is methodological pluralism and why has it been accepted by behavioral scientists?

5. What are the three broad research approaches described in the text? Give an example of the type of question each approach addresses.

6. Describe five of the nine characteristics listed in the textbook that good researchers possess.

CHAPTER 2: *FROM HUNCHES TO TESTABLE HYPOTHESES*

CHAPTER OUTLINE

I. What is Meant by a Cycle of Discovery and Justification?
- A. The philosopher Hans Reichenbach (1938) identified two phases of scientific inquiry:
  1. The **discovery phase** conceptualizes the initial development of research ideas.
  2. In the **justification phase** researchers test their **working hypotheses** and logically defend their conclusions.

II. What Are Hypothesis-Generating Heuristics?
- A. Suitable leads for research can be found everywhere.
  1. McGuire used the term "hypothesis-generating heuristics" to refer to the circumstances or the strategies that were the basis of hypotheses for empirical research. Examples of hypothesis-generating heuristics are:

     a. The effort to make sense of a paradoxical incident.
     b. The use of analogical thinking.
     c. The resolution of conflicting results.
     d. The effort to improve on older ideas.
  2. **Meta-analysis** can be used to develop an overall picture of empirical findings concerning a specific research question as well as an exploratory tool for identifying **moderating variables**.

III. What Is the Potential Role of Serendipity?
- A. Good leads for questions and hypotheses are all around us, and all that is required is to keep our eyes, ears, and minds open.

IV. How Can I Do a Literature Search?
- A. **PsycINFO** is an extensive computerized reference database maintained by the American Psychological Association. This database dates back to 1872 and contains abstracts as well as full-text materials.
- B. Other databases available include census data, full-text data from many scholarly publications (Academic Search Premier, bibliographic records of educational resources (ERIC), news reports by topic areas (LEXIS-NEXIS), full-text dissertations and master's theses (ProQuest Dissertations and Theses), and dictionaries and encyclopedias.
- C. You can consult one of the library's information specialists in your college library for guidance.
- D. PsycARTICLES is another APA database, which is linked with PsycINFO in the libraries that subscribe to both. PsycARTICLES offers full-text articles from all the APA journals, the journals of the Canadian Psychological Association (CPA), and a group of other journals.
- E. It is important to read the actual work, not just the abstract of the work.

V. How Should I Go About Defining Variables?

A. One should begin to think about naming and defining the things one wants to study.

B. **Operational definitions** identify terms on the basis of the empirical conditions used to measure or manipulate them.

C. **Theoretical** (or **conceptual**) **definitions** assign the meaning of terms more abstractly or generally.

D. There are reference sources available that can aid in the development of good operational and theoretical definitions of variables one wishes to study.

VI. What Identifies "Good" Theories and Working Hypotheses?

A. The research idea is molded into a testable supposition, or **working hypothesis** (also called an **experimental hypothesis** in experimental research) based on theory.

B. There is a distinction usually made between hypotheses and theories.

1. A hypothesis is a conjectural statement or supposition.

a. Hypotheses can be derived from a theory.

b. Hypotheses give direction to the researcher's systematic observations.

2. A theory is an organized set of explanatory propositions connected by logical arguments and by explicit and implicit prior assumptions.

a. A theory postulates a kind of conceptual pattern, which can then serve as a logical framework for the interpretation or the larger meaning of one's observations.

b. Seminal theories shape or stimulate other work.

c. Good scientific theories are **generative**, which means they encourage others to generate additional hypotheses.

C. Molding Ideas Into Acceptable Hypotheses

1. A working hypothesis must be plausible, that is, it must have **correspondence with reality** in that it agrees with accepted truths (e.g., other respected theories and reliable empirical data).

2. **Falsifiability** is the most essential criterion for an acceptable hypothesis according to the philosopher Karl Popper. Hypotheses that do not meet this criterion are considered to be outside the realm of science.

3. A hypothesis must be succinct, which is a combination of **coherence** and **parsimony**.

a. **Coherence** refers to whether the hypothesis "sticks together" in a logically compelling way.

b. **Parsimony** refers to how "sparing" or "frugal" the hypothesis is. **Occam's razor** refers to the ruminative and winnowing process of eliminating the superfluous.

VII. What is the Distinction between an Independent Variable and a Dependent Variable?

A. A **variable** is an event or condition that the researcher observes or measures or plans to investigate and that is likely to vary or change.

1. The **dependent variable** is the consequence (or the outcome) in which the researcher is interested.

2. The **independent variable** is the presumed "cause." Changes in this variable lead to changes in the dependent variable.

B. How a variable is labeled always depends on its context.

VIII. What Belongs in My Research Proposal?
A. A proposal might be thought of as a mutual understanding between the student and the instructor.
B. By searching the literature and having discussions with one's instructor, one will be able to develop a rationale for one's hypothesis.
C. A research proposal conveys what one would like to study and how one will go about it.

LECTURE IDEAS AND ACTIVITIES

1. One misconception that students may have concerning the research process is that it is a "boring" endeavor that is relatively straightforward, culminating in a published research report. However, as many researchers know, this is far from the truth. To help students come to appreciate the "exciting" side of the research process, discuss how you became interested in your area of research. Relate the personal process you go through as you formulate and eventually test your research ideas. You may also want to discuss the origin of these research ideas, relating to the various sources of research ideas outlined in the text. Along these lines, there are several books available in which researchers focus on the origins of their ideas and the experiences they had along the way as their research endeavors led them down unexpected and interesting paths.

Brannigan, G. G., & Merrens, M. R. (Eds.). (1992). *The undaunted psychologist: Adventures in research.* New York: McGraw-Hill.
Brannigan, G. G., & Merrens, M. R. (Eds). (1995). *The social psychologist: Research adventures.* New York: McGraw-Hill.
Merrens, M. R., & Brannigan, G. G. (Eds.). (1996). *The developmental psychologists: Research adventures across the lifespan.* New York: McGraw-Hill.

2. To illustrate the point that research ideas can be found almost anywhere, begin a discussion of current events. You may want to read the headlines from that day's newspaper to stimulate this discussion. Have students propose explanations for why these events occurred. After discussing several alternative explanations for the same event, have students shape each of these explanations into plausible research ideas suitable for investigation.

3. The text discusses the role of the literature search in the development of a research proposal. As this may be a student's first exposure to the psychological literature, the student may be unfamiliar with how to critically read a research article. You may want to take this opportunity to discuss the format of the typical journal article, pointing out the purpose of the different sections (i.e., introduction, method, results, and discussion). Appendix A of the text ("Reporting Your Research Results") can serve as a guide for this discussion.

One aspect of reading an empirical article that can be particularly frustrating for students is the statistical analyses the author employed to test the hypothesis. This lack of understanding may lead the students' eyes to "glaze over" when reading the results section, thus missing the major findings of the study. You may want to specifically guide students through a results section of a typical research report explaining how students do not necessarily have to have a sophisticated knowledge of statistics to understand and evaluate the author's major findings and conclusions.

4. Searching the literature can be a daunting task for the beginning researcher, especially for students unfamiliar with the wealth of resources available to behavioral scientists. You may want to discuss how to effectively use library resources to conduct a literature search. Parr (1988) has argued that a general instruction on library usage is a necessary first step in teaching students how to do literature searches. Merriam, LaBaugh, and Butterfield (1999) describe the basic, practical library skills that all psychology students should learn in order to be able to conduct effective literature searches. In addition to the brief discussion of how to find and use reference materials in this chapter, Rosnow and Rosnow (2005) have devoted an entire chapter to this topic in their *Writing Papers in Psychology* manual. Rosnow and Rosnow familiarize students not only with general library operation (e.g., how material is catalogued) but also the types of resources that are often available to students in their literature search.

Merriam, J., LaBaugh, R. T., & Butterfield, N. E. (1999). Library instruction for psychology majors: Minimum training guidelines. In M. E. Ware and C. L. Brewer (Eds.), *Handbook for teaching statistics and research methods (2nd ed.)* (pp. 154–157). Mahwah, NJ: Lawrence Erlbaum.

Parr, V. H. (1996). Course related library instruction for psychology students. In M. E. Ware and D. E. Johnson (Eds.), *Handbook of demonstrations and activities in the teaching of psychology, Vol. 1: Introductory, statistics, research methods, and history* (pp. 132–133). Mahwah, NJ: Lawrence Erlbaum.

Rosnow, R. L., & Rosnow, M. (2005). *Writing papers in psychology (7th ed).* Belmont, CA: Wadsworth.

5. As discussed in this chapter, computerized reference databases represent an easy way to search the literature for relevant references. Feinberg, Drews, and Eynman (1996) have suggested that learning to use these databases may have positive effects on students' attitudes towards the library as well as the literature review process itself (see also Cameron & Hart, 1996). However, using computerized reference databases does have disadvantages which should be discussed with your students (Lewis, 1996). For example, the scope of the database may be limited to articles published only during the past 20–30 years. In addition, the success of one's literature search is influenced by the effectiveness of the search strategy one uses. Parr (1996) describes a general search strategy she uses when working with students who are learning to conduct searches using computerized databases.

Cameron, L., & Hart, J. (1996). Assessment of PsycLIT competence, attitudes, and instructional methods. In M. E. Ware and C. L. Brewer (Eds.), *Handbook for teaching statistics and research methods (2nd ed.)* (pp. 157–161). Mahwah, NJ: Lawrence Erlbaum.

Feinberg, R. A., Drews, D., & Eynman, D. (1996). Positive side effects of online information retrieval. In M. E. Ware and D. E. Johnson (Eds.), *Handbook of demonstrations and activities in the teaching of psychology, Vol.1: Introductory, statistics, research methods, and history* (pp. 136–137). Mahwah, NJ: Lawrence Erlbaum.

Lewis, L. K. (1996). Bibliographic computerized searching in psychology. In M. E. Ware and D. E. Johnson (Eds.), *Handbook of demonstrations and activities in the teaching of psychology, Vol.1: Introductory, statistics, research methods, and history* (pp.138–140). Mahwah, NJ: Lawrence Erlbaum.

6. One way to motivate your students to learn how to effectively conduct library searches is to use "treasure" or "scavenger" hunts. For example, LeUnes (1988) and Mathews (1988) found that students became more comfortable and efficient with doing literature searches after they learned to use the library to answer specific questions as part of a treasure hunt "game." For example, who was the author of the chapter on Personality in the 1971 edition of the *Annual Review of Psychology*? What is the library call number of the *Journal of Psychology*? (Mathews, p. 115) Name eight educational/psychological journals that deal with studies on the development of children. What book did John Watson write in 1928 pertaining to child development? (LeUnes, p. 114)

Gardner (1996) describes another strategy for introducing students to the psychological literature. He provides students with a list of clichés and old sayings (e.g., "opposites attract" or "you can't teach an old dog new tricks") that they are to treat as research hypotheses. The students are then told to find empirical evidence in the psychological literature that either supports or refutes their cliché or saying. Students are required to submit the abstracts from the articles they locate as well as defend why those particular articles would be useful in either supporting or negating the validity of the cliché or saying. This last component of Gardner's exercise is a particularly important aspect of this exercise. While students may become adept at locating articles, they are not necessarily able to evaluate the utility of those articles. By being able to locate the relevant literature and understand how that literature either supports or refutes a research hypothesis, students will gain a better understanding of how to write a clear and focused literature review.

Gardner, L. E. (1996). A relatively painless method of introduction to the psychological literature search. In M. E. Ware and D. E. Johnson (Eds.), *Handbook of demonstrations and activities in the teaching of psychology, Vol.1: Introductory, statistics, research methods, and history* (pp. 129–130). Mahwah, NJ: Lawrence Erlbaum.

LeUnes, A. D. (1988). The developmental psychology library search: Can a nonsense assignment make sense? In M. E. Ware and C. L. Brewer (Eds.), *Handbook for teaching statistics and research methods* (pp. 113–114). Hillsdale, NJ: Lawrence Erlbaum.

Mathews, J. B. (1996). "Hunting" for psychological literature: A methodology for the introductory research course. In M. E. Ware and D. E. Johnson (Eds.), *Handbook of demonstrations and activities in the teaching of psychology, Vol. 1: Introductory,*

statistics, research methods, and history (pp. 131–132). Mahwah, NJ: Lawrence Erlbaum.

7. To help students learn to identify the independent and dependent variables associated with a research question, have them identify the relevant independent and dependent variables in each question below. Then have them provide an operational definition for each.
    1. Do blondes have more fun?
    2. Does taking vitamins increase brain power?
    3. Does age affect how well you can exercise?
    4. Does interacting with relatives cause anxiety?
    5. Does living in close quarters increase the desire to hurt others?
    6. Does exposure to repeated disappointment result in sadness?
    7. Is physical attractiveness related to arrogance?
    8. Do children have more behavioral problems when both parents have careers?
    9. Do patient people take longer to get to the center of a Tootsie Roll Pop?

MULTIPLE-CHOICE QUESTIONS

1. According to the philosopher Hans Reichenbach, it is during the _____ stage of research that scientists formulate the ideas they pursue using the scientific method.
    a. empirical
    b. justification
    * c. discovery
    d. final (20)

2. One has entered the _____ stage of research when scientists actually begin to test their working hypotheses and logically defend their conclusions.
    a. discovery
    b. plausibility
    c. initial
    * d. justification (20)

3. Name one example of a hypothesis-generating heuristic.
    * a. explaining paradoxical incidents in testable ways
    b. identifying the operational definition
    c. falsifiability
    d. the consequence (or the outcome) in which the researcher is interested (22)

4. Irving Janis was fascinated by how a group of intelligent men such as President John F. Kennedy and his advisors could make such a disastrous decision when they authorized the ill-fated Bay of Pigs invasion. To find out, Janis began his study of decision-making in groups that eventually led to the phenomenon of "groupthink." Janis's hypothesis-generating heuristic could best be described as
   a. an attempt to resolve conflicting results.
   * b. an effort to understand a paradoxical incident.
   c. an effort to improve on older ideas.
   d. the use of metaphors. (22)

5. The inspiration for Robert Zajonc's research that ultimately led to his theory of social facilitation was
   a. his desire to improve on older ideas and theories.
   b. his use of analogical thinking.
   c. his trying to make sense of a paradoxical incident.
   * d. his attempt to resolve conflicting results. (23)

6. While studying possible factors related to heart disease, two cardiologists, Meyer Freidman and Ray Rosenman, noticed that there were discernible differences in behavioral patterns between men who were and were not prone to coronary problems. This observation led to the identification of Type A and Type B personality types. The source of Friedman and Rosenman's initial research idea can best be described as
   a. the use of metaphors.
   * b. serendipity.
   c. resolution of conflicting ideas.
   d. improvement on old ideas. (25-26)

7. Databases that are used in psychological research and can be found in the college library include all but which of the following?
   * a. PROinfo
   b. PsycINFO
   c. ProQuest Dissertations and Theses
   d. PsycARTICLES (27)

8. Definitions based on how something will be measured or manipulated are referred to as
   a. theoretical definitions.
   b. conceptual definitions.
   c. precise definitions.
   * d. operational definitions. (29)

9. A researcher defines frustration as the number of teeth marks a student makes on a pencil while trying to solve a difficult problem. This is an example of a(n)
   * a. operational definition.
   b. conceptual definition.
   c. working definition.
   d. theoretical definition. (29)

10. Definitions that use abstract or general terms are considered to be
    a. useful definitions.

b. working definitions.
* c. theoretical definitions.
d. operational definitions. (29)

11. A researcher defines frustration as the negative affect one experiences when trying to solve a difficult problem. This is an example of a(n)
a. operational definition.
b. working definition.
* c. theoretical definition.
d. plausible definition. (29)

12. A testable supposition is also referred to as a
a. theory.
b. construct.
* c. hypothesis.
d. variable. (31)

13. Linda conjectures that as research participants become more frustrated, they will commit more errors on a timed task. Her conjecture is an example of a
a. theory.
b. construct.
c. research idea.
* d. hypothesis. (31)

14. Theories that result in further hypotheses and additional observations are known as _____ theories.
a. coherent
* b. generative
c. parsimonious
d. working (31)

15. Steve hypothesizes that some behaviors are due to one's astrological sign. Steve's hypothesis most likely violates which criterion for acceptable hypotheses?
* a. correspondence with reality
b. falsifiability
c. coherence
d. parsimony (31)

16. Michele hypothesizes that an individual who gets eight hours of sleep, takes a shower, eats a nutritious breakfast, and works under severe time constraints will do worse on a task than an individual who gets eight hours of sleep, takes a shower, eats a nutritious breakfast, but who does not work under the same severe time constraints. Michele's hypothesis violates which criterion for acceptable hypotheses?
    a. correspondence with reality
    b. inclusiveness
  * c. coherence and parsimony
    d. falsifiability (31)

17. Occam's razor is used to address which criterion for acceptable hypotheses?
    a. falsifiability
    b. comprehensiveness
    c. correspondence with reality
  * d. coherence and parsimony (31)

18. A researcher who attributes all successes to his treatment but then attributes all failures to another factor has violated which essential criterion for acceptable hypotheses?
  * a. falsifiability
    b. testability
    c. coherence and parsimony
    d. correspondence with reality (31)

19. An event or condition that a researcher plans to measure or observe is called a
    a. theory.
    b. construct.
  * c. variable.
    d. hypothesis. (31)

20. In an experiment, one manipulates the _____ variable in order to measure its effect on the _____ variable.
    a. dependent; independent
  * b. independent; dependent
    c. theoretical; operational
    d. hypothesis; construct (31-33)

21. _____ is to cause as _____ is to effect.
  * a. Independent variable; dependent variable
    b. Construct; dependent variable
    c. Dependent variable; independent variable
    d. Independent variable; construct (31-32)

22. Bill is interested in how people react to news of an impending snow storm. He informs subjects that it is going to snow 2, 6, or 12 inches and then assesses their anxiety levels. The different snow forecasts represent the _____ variable in this study.
    a. construct
  * b. independent
    c. control
    d. dependent (31-32)

23. A waitress is interested in whether providing candy at the end of a meal to her customers can have a positive impact on the tips she receives. She either does or does not provide candy to her customers when she delivers the final check. The waitress then records the amount of the tip she receives from the dining party. What would the act of providing or not providing candy be in this study?
    a. a construct
  * b. an independent variable
    c. a random variable
    d. a dependent variable (31-32)

24. Bill is interested in how people react to news of an impending snow storm. He informs subjects that it is going to snow 2, 6, or 12 inches and then assesses their anxiety levels. Anxiety levels represent the _____ variable in this study.
  * a. dependent
    b. conceptual
    c. independent
    d. control (31-32)

25. A waitress is interested in whether providing candy at the end of a meal to her customers can have a positive impact on the tips she receives. She either does or does not provide candy to her customers when she delivers the final check. The waitress then records the amount of the tip she receives from the dining party. What would tips be in this study?
    a. a construct
    b. an independent variable
    c. a random variable
  * d. a dependent variable (31-32)

SHORT ESSAY QUESTIONS

1. What is meant by the "discovery" phase of the research process? How is this different from the "justification" phase?

2. Describe three hypothesis-generating heuristics that scientists may use as the basis of hypotheses for empirical research.

3. Describe two strategies one may employ when trying to identify studies relevant to one's research idea.

4. What is the difference between operational and theoretical definitions? Give an example of each.

5. How do hypotheses differ from theories?

6. Describe the three essential criteria for acceptable hypotheses.

7. Why is there no simple classification system for differentiating between variables that are exclusively independent variables and variables that are exclusively dependent variables? Give an example to support your argument.

8. Describe two general categories of independent variables discussed in the text. Provide an example of each.

9. Describe the role of the research proposal in the research process.

CHAPTER 3: *ETHICAL CONSIDERATIONS AND GUIDELINES*

CHAPTER OUTLINE

I. How Do Ethical Guidelines in Research Function?
- A. After the formulation of an appropriate research question, one must contemplate the ethics of the empirical strategy one will use to address the question.
- B. The term **ethics** refers to the values by which people morally evaluate character or behavior.
  1. In science, **ethics** refers to conduct that is considered "morally right" or "morally wrong" as specified by codified and culturally ingrained principles, constraints, rules, and guidelines.
  2. Legal, institutional, and professional **ethical guidelines** contain rules and specifications pertaining to the question, "Should I conduct this study?" when conducting the study involves a moral issue.
  3. The professional guidelines that have figured prominently in the research process were adopted by the American Psychological Association. The APA adopted amendments to the guidelines in 2010.
- C. The General Principles adopted by the APA to address ethical issues centered on the following five basic principles that serve as a framework for the discussion of ethics in this chapter.
  1. Respect for People's Rights and Dignity.
  2. Beneficence and Nonmaleficence.
  3. Justice.
  4. Integrity.
  5. Fidelity and Responsibility.
- D. These five basic principles, while not anticipating every case, remind researchers of their responsibility not to do harm to participants and to do potentially beneficial research in a way that will produce valid results.

II. What Is Informed Consent, and When Is It Used? (Principle I. Respect for Persons and Their Autonomy)
- A. **Autonomy** refers to a prospective participant's right as well as ability "to choose" to take part in the study or to continue in the study.
- B. Researchers have an ethical and legal responsibility to ensure that potential participants know what they are getting into, and that they are free to decide whether or not to participate.
  1. Generally, potential participants indicate this awareness by providing a written agreement (or **informed consent**) to participate.
  2. There are some situations in which obtaining the informed consent of prospective participants is either unnecessary or impossible.
- C. The informed consent form given to prospective participants describes:
  1. The nature of the study.
  2. Any potential risk or inconvenience to them.
  3. The procedure for ensuring the confidentiality of the data.

4. The voluntary nature of their cooperation and their freedom to withdraw at any time without prejudice or consequence.

D. For some studies, the disclosure procedure required for informed consent can become so detailed and cumbersome that it possibly defeats the purpose for which it was intended.
    1. A researcher cannot be said to have complied with the spirit of the law if the informed consent form confuses participants.
    2. In cases where prospective participants are either children or adolescents, the researcher must first obtain consent from a parent or other advocate before approaching the prospective participant.
    3. Participants do not relinquish their legal right to sue the researcher for negligence by signing an informed consent agreement. This right is protected under federal regulations on the use of human subjects (U.S. Department of Health and Human Services, 1983).

III. How Are Ethics Reviews Done and Acted On? (Principle II. Beneficence and Nonmaleficence)
  A. Defining the terms:
    1. **Beneficence** means the "doing of good."
    2. **Nonmaleficence** means "not doing harm."
  B. To satisfy this principle, the researcher submits a proposal of the planned research to a panel of evaluators, called an **institutional review board** (**IRB**), which provides an oversight mechanism by examining the risks and benefits of the proposed study.
    1. Studies classified as **minimal risk** by the IRB are those in which the likelihood and extent of harm to participants is no greater than that typically experienced in everyday life.
    2. Studies that are considered to be minimal risk are usually eligible for an **expedited review**.
  C. A **decision-plane model** may be used to conceptualize the process by which IRBs ideally evaluate the costs and risks of a research proposal.
    1. The risks and benefits of doing a particular study are evaluated on scales of perceived methodological and societal values or interests.
    2. A limitation of this idealized assessment is that it focuses only on the risks and benefits of "doing" research and ignores the societal risks of "not doing" research.

IV. What Are Obstacles to the Rendering of "Full Justice"? (Principle III. Justice)
  A. By justice, one means that the burdens as well as the benefits of research should be distributed fairly.
  B. Justice also implies **fair-mindedness**, or impartiality. However, the notion of impartiality is often one of perception and personal judgment and questions about "what is equal?" or "what is unequal?" are often complex and highly nuanced.

V. How Can a "Relationship of Trust" Be Established? (Principle IV. Trust)
  A. The researcher has an ethical obligation to establish a relationship of trust with the research participants.

B. The assumption is that people will be told what they are getting into (i.e., informed consent) and that nothing will be done to jeopardize this trust.
C. In situations where information must be withheld from participants to prevent possible response biases, the researcher should debrief the participants after their cooperation in the study.
D. One procedure for establishing trust is to protect the **confidentiality** of participants' disclosures. Confidentiality implies that participants' disclosures will be protected against unwarranted access
    1. Ensuring confidentiality seems to lead to more open and honest responding.
    2. A **certificate of confidentiality** is a formal agreement between an investigator and the government agency sponsoring the research that requires the investigator to keep the data confidential and exempts the data from subpoena.

VI. How Do Scientific Quality and Ethical Quality Intertwine? (Principle V. Fidelity and Scientific Integrity)
A. The researcher has an ethical obligation to foster scientific advances that lead to valid knowledge.
B. Poor-quality research is an ethical problem because it is wasteful of resources and can be misleading (even potentially damaging) to society.
C. Table 3.2 describes the relationship between ethical quality and scientific quality. Both are necessary. Principles of scientific quality discussed include:
    1. Transparent – reporting of results is open, frank, and candid.
    2. Informative – enough information is reported to enable sophisticated readers to reach their own independent conclusions and to perform their own calculations.
    3. Precise – results are reported to the degree of exactitude required by the given situation.
    4. Accurate – recording data accurately and not exaggerating results.
    5. Grounded – the methods and statistical procedures are logically and scientifically justified, the questions and hypotheses addressed are appropriate to the design, and the primary data analysis focuses on the questions or hypotheses.

VII. Is Deception in Research Ever Justified?
A. The use of deception is an example of the interrelationship between ethics and scientific integrity.
B. Milgram's experiments sparked ethical debate both inside and outside behavioral and social science.
    1. Diane Baumrind criticized Milgram's research on the basis of how stressful his deception was.
    2. Milgram responded by saying that the chief horror of his experiments was not that stressful deception was carried out, but instead that participants obeyed.
        a. The signs of extreme tension that appeared in some participants were quite unexpected, but his intention was not simply to create anxiety.
        b. Milgram took elaborate precautions to debrief his participants to ensure that they would not feel worse after the experiment than before.

3. Milgram surveyed the participants after they read a full report of the investigation. He found that over 80% of the participants said they were glad to have participated. He regarded the results as providing a moral justification for his research.

C. Adopting a rigid moral orientation that decries deception as wrong would mean banishing all forms of deception or even producing misleading results in some cases.

D. Some argue that deception in any form is morally wrong, whereas others argue that there are special circumstances in which deception is needed to ensure the integrity of important scientific data.

E. Two broad types of deception have been used in behavioral and social research.

1. In **active deception** (or **deception by commission**), subjects are actively misled.
2. In **passive deception** (or **deception by omission**), certain information is withheld from the participants.

F. Ultimately, the decision must be made as to whether a particular potentially harmful deception is worth a possible increase in knowledge.

VIII. What Is the Purpose of Debriefing, and How Is It Done?

A. **Debriefing** provides researchers an opportunity to remove any misconceptions and anxieties the participants may have so that their sense of dignity remains intact and they feel that their time has not been wasted.

1. There are situations in which debriefing may also be impossible or inadvisable.
2. However, in many instances, debriefing is not only ethically essential but can also provide an opportunity to explore what participants thought about the study, providing the researcher with an experiential context in which to interpret the data and with good ideas for further investigation.

B. The following guidelines may be incorporated into the typical debriefing procedure:

1. If the study involved some form of deception, the truth about the research, including why the use of deception was necessary, must be fully explained.
2. Whatever the deception used, participants must be assured that their behavior was due to the effect of the experimental design and does not reflect their intelligence or character.
3. Debriefing should be a gradual and patient process, especially with respect to the details of any deceptions used.
4. **Double deception** should never be used.

IX. How Is Animal Research Governed by Ethical Rules?

A. The use of animals in experiments has been vigorously debated because the very assumption of biological continuities between animals and human beings raises ethical dilemmas, B. Concerns about the treatment of animals has led to federal laws and licensing requirements that articulate the responsibilities of researchers and animal facilities to protect the well-being of experimental animals.

C. Beyond federal regulations, animal researchers are subject to institutional and professional requirements.

D. Despite the heated debate about the use of animals in scientific research, it is clear that society has benefited in terms of biomedical and behavioral advances and that the

ethical consciousness of science and society has been raised with regard to the conduct of animal research.

X. What Ethical Responsibilities Are There When Writing Up Research?
- A. Professional researchers are responsible for making available the data on which their conclusions are based.
- B. It is unethical to misrepresent original research by publishing it in more than one journal and implying that each report represents a different study.
- C. Authors of published articles are expected to give credit where it is due.
- D. Avoiding Plagiarism
  1. The most nagging ethical concern of most instructors is conveying to students the meaning and consequences of **plagiarism** and how to avoid it.
  2. **Accidental plagiarism** occurs when one copies someone else's work but "forgets" to credit it or put it in quotes.
  3. One can use other people's ideas or works in one's research and writing as long as the author of that material is given full credit for originality and one does not misrepresent that material as one's own original work.
  4. Papers saturated with quoted material, while not committing plagiarism, exhibit lazy writing.

LECTURE IDEAS AND ACTIVITIES

1. The current Ethical Principles of Psychologists and Code of Conduct adopted by the American Psychological Association are available online at the American Psychological Association website at http://www.apa.org/ethics/code/index.aspx.

2. The chapter briefly discusses The Belmont Report. The actual report and other federal regulations protecting human research subjects (e.g., Federal Law 45 CRF 46) are available on the Internet at the Office for Human Research Protections website at http://www.hhs.gov/ohrp/index.html. Click on the "policy guidance" link.

3. This chapter lends itself well to the use of structured debates on controversial research practices such as the use of deception or issues of privacy. Rather than focusing on specific issues, you may want students to debate whether controversial studies (e.g., Milgram's obedience studies; Zimbardo's Stanford prison study) should have been conducted. The keys to successful classroom debates are the preparation level of the debate teams as well as the active participation of students who are not on the debate teams. Janet Morahan-Martin (1990) has described how she achieves successful student debates in her introductory psychology classes; these suggestions can be easily adapted to research method classes.

Morahan-Martin, J. (1990). Use of debates in introductory psychology. In V. P. Makosky, C. C. Sileo, L. G. Whittemore, C. P. Landry and M. L. Skutley (Eds.), *Activities handbook for*

the teaching of psychology (Vol. 3, pp. 214–218). Washington, DC: American Psychological Association.

4. Mark McMinn (1999) argues that simply knowing the basic ethical principles one should follow when conducting research is not the same as learning to be an ethical researcher. He suggests that instructors should take a trial-and-error approach to teaching ethics rather than a prescriptive one. To this end, he advocates the use of case studies in which students must take a tentative position as to the ethical nature of the case study before more information about the case study is revealed to them. Through an unfolding process, a simulation of an ethical decision-making process can occur within the classroom such that students can come to appreciate the occasional difficulty scientists can have in making ethical decisions. McMinn has developed a case-study simulation computer program that allows students to take a trial-and-error approach to learning about ethics. This computer program is available from McMinn.

McMinn, M. R. (1999). Ethics case-study simulation: A generic tool for psychology teachers. In M. E. Ware and C. L. Brewer (Eds.), *Handbook for teaching statistics and research methods (2nd ed.)* (pp. 132–133). Mahwah, NJ: Lawrence Erlbaum.

5. Ralph Rosnow (1990; see also Bragger & Freeman, 1999; Strohmetz & Skleder, 1992) has developed a role-playing exercise designed to help sensitize students to the subtleties involved in the ethical evaluations of research studies as well as to illustrate the decision-plane model described in the text. During the first part of this exercise, students are familiarized with the principles of the APA ethics code. Once aware of these principles, students are instructed to find a recent empirical article that, in their opinion, has violated one or more of these ethical principles.

The second step involves the students giving oral reports to the class on the ethical problems with their selected study. During this presentation, students should be able to answer questions about their ethical concerns. After all of the students have discussed their ethical reservations with the studies, students are then asked to become the "author" and to defend their "study" in the face of these ethical concerns. Afterwards, the class is asked to evaluate the costs involved with each study on a scale from 0 to 100 (0 = no moral or ethical cost; 100 = highest ethical or moral cost) as well as the utility of each study on a similar scale (0 = no theoretical or practical utility; 100 = highest theoretical or practical utility). Summarizing the class's ratings of the costs and utility for each study easily leads into a discussion of the decision-plane model described in the text. You may want to plot where each study falls on this decision plane of the costs and benefits of doing the study and have the class discuss whether the true authors were ethically justified in conducting their study based on where the study fell on the decision plane.

Bragger, J. D., & Freeman, M. A. (1999). Using a cost-benefit analysis to teach ethics and statistics. *Teaching of Psychology, 26,* 34–36.

Rosnow, R. L. (1990). Teaching research ethics through role-play and discussion. *Teaching of Psychology, 17,* 179–181.

Strohmetz, D. B., & Skleder, A. A. (1992). The use of role-play in teaching research ethics: A validation study. *Teaching of Psychology, 19,* 106–108.

6. One way to have students consider the importance of each of the five moral principles discussed in the chapter is to have them consider which principle is the most important of the five. That is, if they could adopt only one of the five principles, which principle should that be? Conversely, which of the five principles is the least important? That is, if one wanted to reduce the list down from five principles to four, which principle should be discarded?

7. Robert Brown (1990) has proposed an interesting exercise to help students appreciate the function of an Institutional Review Board. Students are asked to role-play an IRB and evaluate a research proposal that has important theoretical and practical implications but also employs the use of deception in its procedures and incentives. This proposal can be found in the *Activities Handbook for the Teaching of Psychology (Vol. 3)*. After reading the proposal, students are divided into small groups to evaluate the proposal and recommend whether the proposal should be accepted, rejected, or revised for future consideration. Each "committee" should be prepared to justify its decision.

After a discussion of each committee's decision, it is revealed that this proposal was derived from an actual study published in *Science* by Philip Zimbardo and his colleagues (1981). My experience with this exercise is consistent with Brown, who reported that students tend to either outright reject or request revision of the research proposal because of the study's perceived excessive use of deception. Students tend to overlook the theoretical and practical benefits of the proposed study, instead focusing on the perceived excessive costs of doing such research. Reiterating a point made in the text, instructors will want to discuss how ethical decisions are not simple and that researchers must constantly weigh the costs of doing the study against the potential benefits derived from the proposed study.

Brown, R. T. (1990). Ethics of research with human participants. In V. P. Makosky, C. C. Sileo, L. G. Whittemore, C. P. Landry and M. L. Skutley (Eds.), *Activities handbook for the teaching of psychology* (Vol. 3, pp. 249–254). Washington, DC: American Psychological Association.

Zimbardo, P. G., Anderson, S. M., & Kabat, L. G. (1981). Induced hearing deficit generates experimental paranoia. *Science, 212*, 1529–1531.

8. To help students develop a deeper understanding of the ethical implications associated with the use of deception in research, Bernard Beins has employed the Barnum effect to purposely put students into the role of being a deceived subject. He has students complete a bogus personality inventory and then provides each individual an interpretation of his or her results. However, unbeknownst to the students, this interpretation is identical for all students. Students are then asked to evaluate the accuracy of this supposed individualized feedback. After receiving this feedback, Beins informs the class of the ruse and asks them about their feelings at being deceived. Beins reports that his students' initial reaction is to feel gullible and stupid as well as

being mildly distressed. Beins then expands the discussion to focus on how, given their own experiences at being deceived, research participants might feel if they learned they were lied to during the course of an experiment. Through this discussion, students develop a deeper awareness of the effects that the use of deception can have on research participants as well as the need to avoid or minimize the use of deception when designing a research study.

Beins, B. C. (1999). Using the Barnum effect to teach about ethics and deception in research. In M. E. Ware and C. L. Brewer (Eds.), *Handbook for teaching statistics and research methods (2nd ed.)* (pp. 134–136). Mahwah, NJ: Lawrence Erlbaum.

9. Harold Herzog's (1999) article provides an opportunity to extend the text's discussion of the ethical issues concerning the use of animals in research. Herzog briefly reviews two of the basic philosophical positions against the use of animals in research (the utilitarian argument and the rights argument). Herzog then asks his class to assume the role of an "Animal Care and Use Committee" and consider four different research proposals involving the use of animals. Through considerations of these cases, Herzog has found that students developed a deeper awareness of ethical issues surrounding the use of animals in research.

Herzog, H. A. (1999). Discussing animal rights and animal research in the classroom. In M. E. Ware and C. L. Brewer (Eds.), *Handbook for teaching statistics and research methods (2nd ed.)* (pp. 148–153). Mahwah, NJ: Lawrence Erlbaum.

10. One of the challenges students face when learning to write a literature review is how to avoid committing plagiarism. Rosnow and Rosenthal characterize plagiarism as an ethical issue in the research process and devote a section of the chapter to the discussion of plagiarism, including accidental plagiarism and lazy writing (see also Rosnow & Rosnow, 2003). To help students learn how to avoid committing plagiarism, Judith Nye and Allison Scerbo (1994) developed an exercise in which students must identify instances of plagiarism (obvious as well as subtle) in a summary of several paragraphs from a published article. After discussing possible plagiarized passages in the summary, she provides students with one possible summary of the article that avoids these possible plagiarism issues. Handout 3-1 outlines how to conduct this exercise as well as the instructor's key for the plagiarism problems within the summary. Handout 3-2 is the plagiarism exercise that should be provided to all students. Handout 3-3 provides the solution paragraph.

Nye, J. L., & Scerbo, A. (1994). What's wrong with this paragraph? An in-class plagiarism exercise. Unpublished manuscript, Monmouth University.

Rosnow, R. L., & Rosnow, M. (2003). *Writing papers in psychology* (6th ed.). Belmont, CA: Wadsworth.

MULTIPLE-CHOICE QUESTIONS

1. While developing a research proposal, Dr. Smith is debating whether he is justified in deliberately misleading research participants as part of his empirical strategy. Dr. Smith is concerned with the _____ of his research approach.
   a. validity
   * b. morality
   c. adequacy
   d. reliability (41)

2. The desire to maximize the benefits of one's study is central to which ethical principle?
   a. Trust
   b. Nonmaleficence
   * c. Beneficence
   d. Justice (42-43)

3. According to the principle of respect for persons and their autonomy, all individuals have the right to _____ a research study.
   * a. choose whether or not to participate in
   b. recruit other participants for
   c. become a collaborator on
   d. help collect data for (42-43)

4. Which of the following research practices is based on the ethical principle of respect for persons and their autonomy?
   * a. informed consent
   b. debriefing
   c. assurance of confidentiality
   d. minimal use of deception (43)

5. Sometimes informed consent of research participants is not necessary because it would be _____ for the researcher to obtain.
   a. inconvenient
   b. laborious
   * c. impossible
   d. frivolous (44)

6. By signing an informed consent form, an individual has indicated that he or she is willing to _____ the research study.
   a. provide accurate data for
   b. not disclose the hypotheses of
   c. recruit participants for
   * d. participate in (44)

7. Examples of when an informed consent form may not be possible or necessary include all but which of the following?
   a. archival studies that use public records
   * b. studies of young children
   c. risk-free experiments in which instituting informed consent would be counterproductive
   d. use of census data (44-45)

8. An informed consent form typically contains a description of all but one of the following:
   a. the nature of the study
   b. any potential risk or inconvenience to the participant
   c. the procedure for ensuring the confidentiality of the data
   * d. the debriefing procedure (44)

9. Institutional Review Boards are important for complying with which ethical principle?
   a. fidelity and scientific integrity
   b. justice
   c. respect for persons and their autonomy
   * d. beneficence and nonmaleficence (45)

10. Researchers generally submit their planned research proposals to _____ to evaluate the costs (or risks) as well as the benefits of the proposed study.
   * a. an institutional review board
   b. the departmental chair
   c. a colleague from another department
   d. the institution's administration (45)

11. One limitation to how institutional review boards evaluate research proposals is that they may focus mainly on the costs and benefits of _____ research while overlooking the costs of _____ research.
   a. practical; theoretical
   * b. doing; not doing
   c. worthy; unworthy
   d. useless; useful (47)

12. The use of a control group in a treatment effectiveness study can raise ethical concerns based on which principle?
   a. trust
   * b. justice
   c. beneficence and nonmaleficence
   d. fidelity and scientific integrity (48)

13. The ethical principle of justice involves
   a. the use of deception.
   b. invasions of privacy.
   c. violations of confidentiality.
   * d. issues of fair-mindedness. (48)

14. Research participants in Dr. Lane's study are confident that she will maintain the confidentiality of their research disclosures. Which ethical principle is Dr. Lane upholding?

a. respect for persons and their autonomy
* b. trust
c. beneficence and nonmaleficence
d. justice (49)

15. To help participants respond in a more open and honest manner to a survey, the researcher should describe the procedures he will use to ensure the _____ of the participants' responses.
a. reliability
b. utility
c. validity
* d. confidentiality (49)

16. A researcher knowingly and willingly falsifies his data in order to improve the likelihood that his study will be published in a professional journal. Which ethical principle does this violate?
a. trust
b. justice
* c. fidelity and scientific integrity
d. beneficence and nonmaleficence (50)

17. Milgram's experiments are considered controversial because of his use of
* a. deception.
b. confederates.
c. electric shock.
d. authority figures. (51-53)

18. The deliberate misleading of research participants is called _____ deception.
* a. active
b. justified
c. passive
d. unethical (54)

19. Nicole informs research participants that they had either performed well or performed poorly on a simple cognitive test regardless of the participants' actual performance on the test. Nicole is employing the use of _____ deception in her study.
a. justified
b. permissive
* c. active
d. passive (54)

20. The deliberate withholding of information to research participants is called _____ deception.
a. justified
b. permissive
c. active
* d. passive (54)

21. Dan deliberately does not tell his friends that a rumor is unfounded so that he may observe their reactions to the rumor. Dan is utilizing _____ deception in order to study how people react to rumors.
    a. justified
    b. active
    c. unfair
  * d. passive (54)

22. The problem with being completely open and honest to research participants is that
  * a. such honesty may influence the participants' behavior.
    b. it makes the experimenter look foolish.
    c. it raises criticisms from the scientific community.
    d. it violates several ethical principles. (53-54)

23. If a researcher wants to employ the use of deception, the researcher must _____ participants upon completion of the study.
    a. pay
  * b. debrief
    c. console
    d. reassure (55)

24. Revealing the true nature of the research study as well as the necessity for the use of deception is part of a procedure known as
    a. elaboration.
  * b. debriefing.
    c. a cost-benefits analysis.
    d. ethical evaluation. (55)

25. Which of the following should a researcher NEVER do when debriefing?
    a. Eliminate possible feelings of gullibility in the participants.
    b. Emphasize the role of participants as "co-investigators" in the research venture.
  * c. Use double deception as part of the research design.
    d. Provide an explanation for the necessity for the use of deception. (55)

26. The rationale for using animals in research is based on the assumption that
    a. the ethical requirements are more lenient for research involving animals as compared to humans.
    b. human participant research is more expensive than animal research.
    c. research involving animals produces more valid results than research involving humans.
  * d. there are biological continuities between humans and animals. (56)

27. Issues of plagiarism are relevant to which ethical principle?
    a. trust
  * b. fidelity and scientific integrity
    c. beneficence and nonmaleficence
    d. justice (57)

SHORT ESSAY QUESTIONS

1. Why do ethical considerations play an important role in the research process?

2. How can a researcher satisfy the ethical principle of respect for persons and their autonomy?

3. Differentiate between beneficence and nonmaleficence. Why are both important considerations in the research process?

4. Describe how an institutional review board ideally should evaluate a research proposal that has been submitted for ethical approval.

5. Why should a researcher consider not only the ethical issues involved with conducting a study but also the ethical issues involved with not conducting that study?

6. Why is deception sometimes employed in research involving humans? Describe the two ways that a researcher may deceive research participants.

7. Why is it necessary to debrief research participants? How should this debriefing be conducted when the use of deception has occurred?

8. What is meant by “plagiarism”? Why is it relevant to ethical considerations in the behavioral sciences?

Handout 3-1

WHAT'S WRONG WITH THIS PARAGRAPH?

Instructor's Key

Steps to follow while doing this exercise:
1. Give out writing exercise handout.
2. Students work individually to identify the problems in the writing sample.
3. Students work in groups to fix these problems.
4. Groups show their solutions to instructor.
5. Instructor hands out solution page.

Key For Problem Paragraph:
Underlined: plagiarism problems that must be corrected.
Bold: not plagiarism, but a poor choice for quotation; should be paraphrased instead.

Compliments are rewarding and fun, but they have their bad side too. According to Knapp, Hopper, and Bell (1985), the difficulties and contradictions associated with compliments make them peculiarly fascinating to us. They are a threat that we nonetheless crave. On the one hand, we desire compliments because they make us feel good about ourselves. On the other hand, they often put us on the defensive or raise our skeptical antennae because they come across as insincere. "**The frequency with which compliments are used in our culture increases the problem of believability**" (p. 25). Such uncertainty about the sincerity of compliments can cause us discomfort because we do not know how to deal with them. For example, **Knapp et al.** found that 66% of their subjects reported feeling uncomfortable, defensive, or cynical after receiving compliments. Still, compliments can have powerful and positive effects on our personal and vocational lives—we love to get them.

Handout 3-2

WHAT'S WRONG WITH THIS PARAGRAPH?

Writing Exercise

The following paragraph has many things wrong with it. The author has borrowed too much from the original authors' writing, failing to paraphrase completely and documenting quotations inadequately. There is also one case of inaccurate citation of a source. Your job is to track down the mistakes and correct them. Save this paragraph from the plagiarism police!

Compliments are rewarding and fun, but they have their bad side too. According to Knapp, Hopper, and Bell (1985), the difficulties and contradictions associated with compliments make them peculiarly fascinating to us. They are a threat that we nonetheless crave. On the one hand, we desire compliments because they make us feel good about ourselves. On the other hand, they often put us on the defensive or raise our skeptical antennae because they come across as insincere. "The frequency with which compliments are used in our culture increases the problem of believability" (p. 25). Such uncertainty about the sincerity of compliments can cause us discomfort because we do not know how to deal with them. For example, Knapp et al. found that 66% of their subjects reported feeling uncomfortable, defensive, or cynical after receiving compliments. Still, compliments can have powerful and positive effects on our personal and vocational lives—we love to get them.

Below is the original source for the material used in the plagiarized paragraph.

Compliments—we love to get them. They let us know we are appreciated; they make us feel good about ourselves. And we love to give them. They make people like us; they make people do things for us. But, compliments have their bad side too. We may react negatively if we think a compliment has been delivered insincerely or that we are being set up to do something we don't want to do. We usually don't appreciate getting praise from someone we think lacks intelligence or taste. And the frequency with which compliments are used in our culture increases the problem of believability. Most letters of recommendation, for example, are overwhelmingly complimentary, which both cheapens the praise and makes it difficult to decide what is true and untrue. Even when praise is sincere, we may have difficulty in knowing how to respond. Sociologists Ronny Turner and Charles Edgley reported that two-thirds of the 245 people they observed receiving compliments later said they felt uncomfortable, defensive or cynical about the compliment. The difficulties and contradictions associated with compliments make them peculiarly fascinating. They are a potentially threatening phenomenon that people seem to crave; a form of behavior that has powerful and positive effects on our personal and vocational lives, despite its suspect credibility; and an aspect of conversation that people experience every day yet still have trouble dealing with.

Knapp, M. L., Hopper, R., & Bell, R. A. (1985, August). I really loved your article, but you missed your deadline. *Psychology Today, 19,* 25–28.

Handout 3-3

WHAT'S WRONG WITH THIS PARAGRAPH?
Solution Paragraph

Congratulations! You've taken the first steps in correcting plagiarism problems common to student writing. To be sure you've been thorough in rooting out the plagiarism, the following paragraph shows one solution to the problems portrayed in the original paragraph.

Compliments are rewarding and fun, but can cause confusion or leave us wondering about their authenticity. According to Knapp, Hopper, and Bell (1985, p. 26), "the difficulties and contradictions associated with compliments make them peculiarly fascinating" to us. On the one hand, we desire compliments because they can improve our self-esteem; on the other hand, they often raise our skeptical antennae because we question their sincerity. We are especially concerned about the integrity of compliments when we suspect there are ulterior motives behind them. Such uncertainty can be embarrassing and awkward. For example, Turner and Edgley found that 66% of their subjects reported feeling wary and distrusting after being complimented. Still, compliments can sometimes make all the difference to us. In short, they are a personal threat that we nonetheless long for (Knapp et al., 1985).

CHAPTER 4: *METHODS OF SYSTEMATIC OBSERVATION*

CHAPTER OUTLINE

I. What Is Meant By Systemic Observation?
- A. **Systematic observation** means that what the researcher observes, and how it is recorded, uses a methodology that can be evaluated on the basis of technical standards.
  1. Given the confining or limiting nature of any single observational method, scientists often employ multiple methods of observation (or **methodological triangulation**) to obtain a more coherent view of the pattern of interest.
  2. These observations are, at least in part, guided by certain preexisting plans, questions, or hypotheses.
- B. A common distinction is made between qualitative and quantitative research.
  1. **Quantitative research** encompasses procedures and techniques in which the observed data are recorded in numerical form.
  2. **Qualitative research** refers to procedures and techniques for collecting data that exists in other than a numerical form, such as narrative or pictorial.

II. How Do Researchers Simultaneously Participate and Observe?
- A. **Participant observation** refers to studies in which a group or community is studied from within by a researcher who records behavior as it occurs.
- B. Advantages of participant observer research:
  1. It allows one to record events as they occur rather than relying on public records of past events.
  2. It enables one to watch events in their "wholeness," particularly those that would be impossible to simulate in a lab, or which might be too sensitive or too risky to try to manipulate experimentally.
- C. While participant observers attempt to approach the field of observation without any preconceived ideas, some theoretical preconception is necessary for the observer to know what, where, when, or how to study what they are interested in.
- D. Participant observers usually work in teams to avoid observer bias.

III. What Can Be Learned From Quantifying Observations?
- A. Sociometric methods have developed into Social Network Analysis (SNA), a means of mapping interactions and relationships. SNA can be used to graph pathways of interpersonal behavior in social networks.
- B. A term for those who control the flow of information is a gatekeeper. Cult leaders and CEOs are examples of gatekeepers.

IV. How Are Judgment Studies Done?
- A. Coders or raters can be selected in one of three ways:
  1. On the basis of intuition. (informal)
  2. By consulting the research literature for relevant criteria to help them choose. (formal)

3. By doing **pilot testing** with the people who volunteered to participate as raters to assess their accuracy of judgment on the relevant issues.

B. Judges or raters are used in judgment studies.
   1. Allow measure of judge-to-judge reliability.
   2. Allow use of basic statistics to summarize data.

V. How Does Content Analysis Work?

A. The use of **archival material** is an example of a category of observation known as **secondary observation**.

B. A popular method for studying written messages and pictorial documents is called **content analysis**.

C. There are three general guidelines for doing a content analysis.
   1. The analysis of content should be consistent among the judges. That is, there should be satisfactorily high **judge-to-judge reliability**.
   2. The specific categories and units should be relevant to the questions or hypotheses of the study.
   3. It is important to develop a good sampling procedure.

D. Content analysis is limited by the quality, dependability, and relevance of the material to be analyzed.

E. However, there are four definite advantages of content analysis when used properly.
   1. Developing a coding system and then implementing it requires little more than common sense logic.
   2. It is a "shoestring" methodology in that, although labor intensive, it does not require much capital investment.
   3. It is a "safe" methodology, because the researcher can add necessary information if it is missed or incorrectly coded.
   4. It forces researchers to scrutinize the material that they are evaluating and classifying.

VI. How Are Situations Simulated in Controlled Settings?

A. In human experimental research, it is frequently possible to simulate a causal relationship in a controlled experimental setting in which one can manipulate the causal condition.

B. While doing research in a laboratory setting is a convenient and effective way of studying a phenomenon of interest, caution is warranted with generalizing from laboratory simulations to situations outside the lab.

C. Efforts to improve the realism and generalizability of simulations have led to additional innovations and virtual-reality technology to use in the lab.

VII. What Are Plausible Rival Hypotheses and the Third-Variable Problem?

A. One should think carefully about **rival interpretations** or **rival hypotheses** when reading the results of studies.

B. Plausible rival hypotheses should be considered so that, as the research process progresses, alternative hypotheses are considered and evaluated as to the degree of confirmation they confer on a theoretical explanation. The idea is that the fewer the

plausible rival hypotheses remaining, the greater the likelihood of confirming the remaining interpretations.

C. The researcher must also consider the third-variable problem, which is the idea that another variable that is correlated with the variables under study and impacts both of them. If two variables, A and B, appear to be related, a researcher should also consider that another variable, C, is responsible for the observed relationship between A and B.

VIII. What Is the Distinction Between Reactive and Nonreactive Observations?

A. **Reactive observations** are observations that affect the behavior being observed.

B. **Nonreactive observations** are observations that do not affect the behavior being observed.

C. **Concealed measurements** and **partial concealment** are examples of nonreactive observations.

D. **Unobtrusive observation** involves the use of concealment such that the persons being studied are unaware that they are being observed for the purpose of research.

1. 1. Eugene J. Webb and his colleagues have classified unobtrusive measures into four broad categories:

a. **Archival Records**

b. **Physical Traces**

c. **Simple Observations**

d. **Contrived Observations**

IX. A Final Note

A. Scientists, like all human beings, are susceptible to the biases imposed by limitations of perception and cognition.

B. Independent replications are therefore important as a way of checking on the accuracy of any single observation or set of observations.

LECTURE IDEAS AND ACTIVITIES

1. Paul Woods (1981) describes a simple exercise that demonstrates the advantage of using systematic rather than casual, everyday observations when engaging in scientific inquiry. Woods has his class led on a short walk around campus by either a colleague or a class member. The students are instructed to carefully observe what is happening as they walk around and to be prepared to answer questions about their observations upon the completion of the walk. As the students take this walk, Woods follows closely behind, recording approximately 50 factual questions and their answers on a portable tape recorder based on his own observations made during that walk (e.g., “How many planes flew overhead? What was the first person we passed doing?”). Upon completion of the short walk, Woods plays his recorded questions and the students write down their answers. Woods then tabulates the number of right and wrong answers to each question. Woods found that this exercise easily leads into a discussion for the need of developing a structured method of making observations and the need to specify what is to be observed before one begins collecting observations. In inclement weather, this activity may still

be done by having students go to a large public area such as the student center and observing the activities that are occurring around them.

Woods, P. J. (1981). Accuracy of observation. In L. T. Benjamin, Jr. and K. D. Lowman (Eds.), *Activities handbook for the teaching of psychology* (Vol. 1, pp. 5–6). Washington, DC: American Psychological Association.

2. One of the challenges an instructor may face is helping students to understand the complexities associated with conducting even simple observational studies. To this end Bernard Beins has developed an activity to demonstrate some of the problems that can occur when engaging in the participation observation approach. Beins recruits two volunteers from his class. Taking these students outside of the classroom, he explains to them that they will be acting as observers of the class. Their job will be to record the number of fidgets they observe occurring among the other students in the class during the class activity. Beins prevents the student-observers from asking any questions about the instructions (e.g., how should one define a "fidget"?) as these questions will become part of the discussion after completion of the activity.

Beins and the student-observers return to the classroom, whereupon Beins begins a series of five one-minute lecture segments during which the student-observers are to record the number of fidgets they observe among their fellow classmates. During the first segment, Beins discusses a topic unrelated to systematic observation to provide a baseline period for the number of fidgets among the students. During the second one-minute segment, Beins instructs the students to close their eyes and imagine that insects are crawling on their skin. In the third segment, the students are sitting with their eyes closed, imagining that insects are crawling on them (this allows the students to imagine the insects without other distractions, thus increasing the likelihood of fidgets occurring). During the fourth and fifth one-minute segments, Beins asks the class to speculate about the purpose of the activity as well as the role of the student observers. These last two segments serve as a "cool down" period to provide a "post-insect" baseline.

Upon completion of the activity, Beins has the student-observers share the number of fidget observations they made during each of the one-minute segments. Very quickly it will become obvious that there are noticeable differences in the observers' counts. Beins uses these differences as a springboard for a class discussion as to the difficulties associated with using naturalistic observations as well as how these difficulties may be overcome. Beins notes several problems that naturalistic observations may pose. For example, observers may differ in how they are operationally defining what they are observing, thus reinforcing the importance of concrete definitions of variables of interest. Also, observers may differ in how they are recording the data (e.g., recording immediately every occurrence of the behavior or keeping a mental tally of the occurrences, which is only actually recorded at the end of each observation period). Beins discusses how even differences in one's vantage point for observing the activities of interest can lead to differences in reported occurrences of the behaviors. Finally, Beins asks students about the possible effects of being aware that one is being observed may have on one's behavior (i.e., the reactivity problem). Beins has found that this five-minute activity helps students recognize problems and limitations that are inherent even in "simple" naturalistic observations.

Beins, B. C. (1999). Counting fidgets: Teaching the complexity of naturalistic observation. In L. T. Benjamin, B. F. Nodine, R. M. Ernst and C. B. Broeker (Eds.), *Activities handbook for the teaching of psychology* (Vol. 4, pp. 53–55). Washington, DC: American Psychological Association.

3. Andrea Zeren and Vivian Makosky (1988) describe an activity where they utilize television programs to provide students with an in-class opportunity to make systematic observations of human behavior. While Zeren and Makosky specifically focus on the observational methods of time sampling, event sampling, and trait ratings, this exercise can be easily adapted to demonstrate the observational techniques discussed in the text.

Zeren and Makosky begin the exercise with a lecture on systematic observational methods in general, and then they focus on the structured observational methods of time sampling, event sampling, and trait ratings. For the next lecture, the class is divided into smaller groups of approximately 3–5 students. Each group is responsible for developing all of the components for one of the three observational methods. Once each subgroup has finalized its observational strategy, the class as a whole watches a half-hour videotaped television program. The students independently record their own observations, using the observational strategy developed in their subgroup. Upon completion of the television program, Zeren and Makosky have each group compute the interrater reliability among all of the subgroup members. Finally, each subgroup is instructed to discuss the advantages and disadvantages of their method of observation as well as how any shortcomings may be overcome.

Zeren, A. S., & Makosky, V. P. (1988). Teaching observational methods: Time sampling, event sampling, and trait rating techniques. In M. E. Ware and C. L. Brewer (Eds.), *Handbook for teaching statistics and research methods* (pp. 144–146). Hillsdale, NJ: Lawrence Erlbaum.

4. Anita Meehan (1993) has developed a simple but effective demonstration of how interpreter effects can bias observations. Before the beginning of class, Meehan secretly places a cigarette with a white filter on the chalk tray. After discussing the importance of empirical observations, Meehan challenges the class's observational skills. She goes to the chalk tray, picks up the cigarette and, while holding it up, asks the class to describe the object they are observing. Students frequently begin with responses such as "It's white," and quickly move on to inferences, such as "It's chalk" or "It's used to write on a chalkboard." Students are often surprised to find out that what they are observing is a cigarette, not a piece of chalk. This simple demonstration easily leads into a discussion of what observers can do to minimize such interpreter effects when making observations. This demonstration may also be applicable when discussing experimenter expectancy effects. This exercise can be successfully adapted for use in classrooms that use whiteboards rather than chalkboards by having the instructor pull the cigarette out of a box of chalk.

Meehan, A. M. (1993, October). *Pure observations are hard to come by.* Paper presented at the Teaching of Psychology Conference, Ursinus College, Collegeville, PA.

5. To provide students with experience conducting ethnomethodological research, have them study a college athletic event, such as a football or basketball game, and the activities associated with it. Students may want to use the generic ethnographic questions described in the class. Afterwards, the students can discuss their findings as they try to "make sense" of why such athletic events are popular among the student body and why they and other students have emotional loyalties to their college teams.

6. Fontes and Piercy (2000) have developed a series of activities they use to demonstrate the utility of qualitative research approaches when addressing a scientific question of interest. One of the exercises they have developed is intended to introduce students to ethnomethodology. Fontes and Piercy point out to their students that because ethnomethodologists are interested in ordinary, routine, everyday behavior within a society, they will often purposely disrupt this routine by doing something out of the ordinary. Through this disruption, the society's unspoken norms of behavior become more evident. Fontes and Piercy instruct their students to engage in an "out of the ordinary" behavior in order to observe the reactions of others to this behavior. By considering these reactions, the students are then asked to write a short paper in which they discuss what tacit norms may be governing individuals' behaviors in that situation. Fontes and Piercy provide some suggested behaviors students may engage in (e.g., while in an elevator, standing with one's back to the door; while standing in line, shaking hands while introducing oneself to the other individuals also standing in line). While students are free to propose the out of the ordinary behavior they wish to engage in, Fontes and Piercy emphasize that the students should discuss their proposed experiment with either the instructor or the rest of the class to minimize inappropriate or ethically questionable (see Chapter 3) experiments. Fontes and Piercy find that by engaging in norm violation experiments, students become more aware of the value of ethnomethodology as a research approach.

Fontes, L. A., & Piercy, F. P. (2000). Engaging students in qualitative research through experiential class activities. *Teaching of Psychology, 27,* 174–179.

7. Angela Lipsitz (2000) incorporates the use of archival research in a classroom exercise to introduce topics concerning research methodology. Lipsitz has students investigate the research question of whether there are gender differences with respect to smiling by examining individual yearbook photos. Each student is instructed to bring a yearbook from a coed school to class. After developing an operational definition of "smiling," the students examine the yearbook photos from one class (e.g., juniors), noting whether or not the person was smiling and the person's gender. At the end of class, Lipsitz records this information (in percentage form) from each student on the board or overhead transparency, as well as the yearbook publication date, the type of school (e.g., high school), the class year observed, the total class size, and the coder's gender. By examining the pattern of results, the class is asked to draw conclusions concerning whether there is a gender difference with respect to smiling. Lipsitz has found, consistent with prior research, a greater percentage of the women are typically found to be smiling as compared

to the men. Finally, Lipsitz uses this opportunity to discuss the methodological and theoretical limitations of their study.

Lipsitz, A. (2000). Research methods with a smile: A gender difference exercise that teaches methodology. *Teaching of Psychology, 27,* 111–113.

8. To demonstrate the utility of content analysis for answering a researcher's questions, you may want to have the students conduct their own content analysis. For example, you may want to pose the question to the class, "Does the outcome of a local sporting event determine how much coverage it is given by the newspaper?" Students can then sample articles from local newspapers that report the outcomes of sporting events and code these articles in a manner that addresses the instructor's question. For example, students may decide to code the articles with respect to article length or the adjectives used in the article. As the class conducts this content analysis, the instructor can discuss the advantages and disadvantages of using this methodology.

9. Sandra Carpenter (1998) describes a content analysis project that is appropriate for students with minimal exposure to research methodology. Carpenter has her class use content analysis to investigate whether societal stereotypes of a group are evident in how members of that group are portrayed in the media. To make the project manageable to even "research novices," she breaks it up into four stages. During the first stage, students are instructed to find a journal article on stereotypes in the media so that they can develop a realistic hypothesis to test. In the second stage, students must develop the method they will use to test the hypothesis. More specifically, students must decide which medium they will use, how they will sample from this medium, and the coding categories they will use in their content analysis. In the third stage, the students submit their proposal for data collection to the professor for review. At this stage, students must also develop a strategy for analyzing the data (which usually involves simple descriptive statistics). Once the data collection and analysis has been completed, students enter the fourth stage. Carpenter has students consider theoretical issues concerning the external validity of their project as well as the implications of their results. Upon completion of the entire project, Carpenter has her students share their research findings with each other, usually in the form of a poster presentation (although this presentation can easily be done in a written or oral format). Carpenter notes that by breaking this content analysis project into smaller, more manageable tasks for students, this project can easily be completed even within a restricted time span, as is often the case with summer or short-term courses.

Carpenter, S. (1998). Content analysis project for research novices. *Teaching of Psychology, 25,* 42–44.

MULTIPLE-CHOICE QUESTIONS

1. Using multiple methods of observation to "zero in" on the effect of interest is called methodological
   a. robustness.
   b. scrutiny.
   * c. triangulation.
   d. ecumenism. (63)

2. Observation that is guided by a particular plan or involves a system that can be evaluated on the basis of technical standards is referred to as _____ observation.
   a. everyday
   b. pure
   c. unbiased
   * d. systematic (63)

3. Research that involves the collection of data that is in a nonnumerical form is referred to as _____ research.
   a. empirical
   * b. qualitative
   c. quantitative
   d. systematic (63)

4. John is interested in how older individuals cope with institutional living, so he is visiting a nursing home and tape recording conversations with some of the residents. His research methodology would be considered to be
   a. quantitative.
   * b. qualitative.
   c. sufficient.
   d. contrived. (63)

5. Research that involves the recording of data in numerical form is referred to as _____ research.
   a. naturalistic
   b. qualitative
   c. robust
   * d. quantitative (63)

6. While observing individuals living in a nursing home, John counts the number of times a resident must ask for help before help is provided. John is making _____ observations.
   * a. quantitative
   b. accurate
   c. qualitative
   d. reliable (63)

7. Numerical observations are referred to as _____ observations while nonnumerical observations are called _____ observations.
   * a. quantitative; qualitative

b. systematic; nonsystematic
c. nonsystematic; systematic
d. qualitative; quantitative (63)

8. To understand the role of camaraderie on a sports team, Kim rides on the team bus so that she can observe the players' interactions before and after every game. Kim's research is an example of
a. field experimentation research.
b. partial concealment research.
* c. participant observer research.
d. secondary observation research. (64)

9. The use of archival material is an example of _____ observations.
a. primary
* b. secondary
c. ethnographic
d. participant (72)

10. Judy is interested in whether a local newspaper tends to have a conservative or liberal constituency of readers. She randomly selects letters to the editor from the paper's op-ed section and then codes the letters based on whether they express conservative or liberal ideas. Judy is trying to answer her question by doing a(n)
a. field experiment.
b. ethnography.
c. laboratory experiment.
* d. content analysis. (72)

11. To examine the use of politeness among children, Kim records conversations between school children and their teachers. She then randomly samples some of these conversations and counts the number of times the children said "please" or "thank you" during the conversation. Which research methodology is Kim employing to answer her research question?
a. participant observation
* b. content analysis
c. ethnography
d. experimentation (72)

12. Which of the following is NOT a general guideline noted in the text for successfully doing a content analysis?
a. One should properly train the judges to improve judge-to-judge reliability.
b. One should develop specific categories and units of analysis.
* c. One should select more intelligent judges to evaluate the archival material.
d. One should develop a good sampling procedure for selecting the archival material. (73)

13. All of the following are possible strategies for selecting independent judges to assist in making research observations EXCEPT
a. the use of pilot testing.
* b. the potential judges' understanding of the research hypotheses.
c. the reliance on the researcher's intuitive sense of what type of judges one should use.

d. the consultation of the literature for identifying what type of judges one should use. (69)

14. The principal benefit of a laboratory experiment is that it allows a scientist to simulate _____ in a highly controlled setting.
* a. causal relationships
b. underlying behaviors
c. real-life situations
d. controversial issues (73)

15. Alternative explanations for the results of a study are called
a. complementary explanations.
* b. rival hypotheses.
c. hypotheses of interest.
d. research problems. (75)

16. In a study of cancer, it was observed that higher levels of milk consumption were positively correlated with higher incidence of cancer. Instead of claiming that higher milk consumption directly leads to higher incidence of cancer, the researcher explained that this correlation was explained by increased longevity. This is an example of
* a. the third-variable problem.
b. rater to rater reliability.
c. reactive observations.
d. a judgment study. (75-77)

17. The use of some form of concealment is characteristic of _____ observations.
a. unethical
* b. nonreactive
c. participant
d. reactive (77)

18. Observations that affect the behavior of those being observed are called _____ observations.
a. partially concealed
b. concealed
* c. reactive
d. nonreactive (77)

19. Michelle interviews shoppers at a local food store about their food preferences, but she is really noting their use of food coupons. Michelle is employing the use of
a. concealment.
* b. partial concealment.
c. field experimentation.
d. secondary observations. (77)

20. Observations made by the researcher without the participants' awareness are called _____ observations.
a. systematic
* b. unobtrusive
c. participant
d. secondary (78)

21. All but one of the following are types of unobtrusive observations:
a. physical traces.
b. simple observation.
c. archival records.
* d. mundane realism. (78)

22. Measuring tread wear on a carpet in order to investigate foot traffic patterns is an example of which type of unobtrusive measure?
a. simple observations
b. contrived observations
c. archival records
* d. physical traces (78)

23. Chris sits at a local mall and watches how adolescents interact with each other. This is an example of which type of unobtrusive observation?
a. contrived observations
* b. simple observations
c. physical traces
d. archival records (78)

24. To investigate whether the enrollment of nontraditional students has increased at her college, Diane examines the ages noted on student applications for the past ten years. This is an example of which type of unobtrusive observation?
a. simple observations
* b. archival records
c. physical traces
d. contrived observations (78)

25. Interested in how people would react to someone who deviates from the norm, Tom has a friend go to class in a bathrobe and then observes how people react to his friend. This is an example of which type of unobtrusive observation?
    a. physical traces
    b. archival records
    c. simple observations
  * d. contrived observations (78)

SHORT ESSAY QUESTIONS

1. Why must scientists rely on systematic observations rather than casual, everyday observations to test their research ideas?

2. What is meant by "methodological triangulation"? Why is it important in psychological science?

3. Differentiate between quantitative research and qualitative research. Give an example of each.

4. What is participant observation? Provide an example of how a researcher might employ participant observation in order to answer a research question of interest.

5. Why is the use of archival material considered to be a secondary observation? Provide an example of how archival material might be used to address a research question of interest.

6. Describe the three general guidelines for doing a content analysis.

7. Why can experimental simulations help to identify causal relationships?

8. What is a rival hypothesis? Why are rival hypotheses of such concern to researchers?

9. What is the third-variable problem? Provide an example.

10. What is the difference between reactive and nonreactive observations? Give an example of each.

11. What is the difference between concealed measurements and partial concealment measurements? Provide an example of each.

12. Describe the broad categories of unobtrusive observations identified by Eugene J. Webb and his colleagues.

13. Describe two approaches that a researcher may use when deciding on the types of judges needed for an observation study.

CHAPTER 5: *METHODS FOR LOOKING WITHIN OURSELVES*

CHAPTER OUTLINE

I. What Are the Uses and Limitations of Self-Report Measures?
- A. **Self-report measures** involve individuals looking within themselves and describing their attitudes, feelings, perceptions, and beliefs.
- B. **Standardized measures** are self-report measures that were developed and are administered and scored according to certain rules or standards.
- C. Many behavioral and social researchers use a variety of self-report measures in their work.

II. Four Important Issues
- A. How dependable is the self-report data?
  1. One must make the assumption that what research participants report is true and not merely a strategy to make the participants "look good."
  2. **Evaluation apprehension** may lead people to be evasive or not completely forthcoming, particularly when the questions being asked are sensitive in nature.
  3. Assuring respondents that their responses will be held in strict confidence as well as allowing respondents to answer privately may reduce evaluation apprehension.
- B. What are the ethical implications of using self-report methods? For example, research participants have the right to withhold information and the right not to have the information they disclose made public or used against them.
- C. Can the information provided by self-reports be regarded as valid and as reliable as other behavioral data?
- D. Interpretation of individual scores is problematic if the measure is not norm referenced.
  1. Single scores between two individuals may not be comparable.
  2. Pre-post differences are inherently meaningful, as are between-group differences with random assignment.

III. What Are Open-Ended and Fixed-Choice Items?
- A. **Open-ended** questions allow one the opportunity to express one's feelings and impressions spontaneously.
  1. Advantages of open-ended measures are that:
     a. They do not lead the respondent by suggesting specific answers.
     b. Their approach is exploratory, allowing the researcher to find out whether the person has anything at all to say.
     c. They invite the person to answer in his or her own language, a procedure that can sometimes help to increase rapport.
  2. Disadvantages of open-ended questions are that:
     a. They are time consuming for the researcher.

b. They often elicit rambling and off-the-mark responses that may never actually touch on the topic of interest.
c. They may be hard to assess for reliability.

B. **Fixed-choice measures** use a more controlled approach, giving respondents specific options such as yes/no or multiple choice alternatives.
1. The advantages and disadvantages of fixed-choice measures are essentially the reverse of those for open-ended measures.
2. The major advantage of the fixed-choice format is that it forces respondents' replies into the dimensions of interest to the researcher rather than producing irrelevant or uncodable answers.

C. The rule of thumb for deciding whether to use open-ended or fixed-choice measures is that the measures chosen should match the dimensions of interest as well as the kind of information desired.

IV. How Are Personality and Projective Tests Used?

A. One of the oldest, but still frequently used, self-report personality measures is the **projective test**. Projective tests use an open-ended format.

B. The **Rorschach test** is an open-ended measure that instructs respondents to describe whatever they see in an inkblot. Professionally supervised experience is required to learn how to score and interpret the Rorschach.

C. The **Thematic Apperception Test** (TAT) is an open-ended measure that asks respondents to make up a story explaining the events occurring in a picture. The stories are presumed to disclose perceptions of interpersonal relationships.

D. A well-known personality measure with a fixed-choice format is the **Minnesota Multiphasic Personality Inventory** (MMPI).

V. What Is Meant By Measuring Implicit Attitudes?

A. Attitude is one of the core constructs in the field of social psychology.

B. Attitude is related to opinion with the distinction that opinions are verbal entities that can be measured directly, whereas attitudes are inferred.

C. Attitude questionnaires are a typical way to collect verbal responses (opinions) on issues that presumably reveal the person's underlying attitude.

D. An alternative approach to the attitude questionnaire are tests to measure implicit attitude through the participant's actions or judgments through automatically activated evaluation, without the performer's awareness of the cause. The Implicit Attitude Test (IAT) developed by Greenwalk and Banaji is an example of an implicit attitude measure.

VI. What Are Numerical, Forced-Choice, and Graphic Ratings?

A. **Numerical scales** are rating scales in which respondents work with a sequence of defined numbers. These numbers may be explicitly stated for the respondent to see and use, or they may be implicit.

B. **Forced-choice scales** "push" or "force" a respondent into making a definite rather than a neutral statement by having the respondent choose between equally favorable (or equally unfavorable) response alternatives. This type of scale is intended to minimize the response bias called the **halo effect**.

1. The **halo effect** occurs when the person doing the rating forms a very favorable impression of the target person based on one trait and extends that impression to the person's other traits.

C. **Graphic scales** (thermometer scales) require participants to respond to a question by making a mark along a straight line that is anchored on each end by **bipolar** items, or extreme opposites.

D. **Segmented graphic scales** divide the straight line into segments, transforming the scale into a numerical rating scale.

VII. What Are Rating Errors and How Are They Controlled?

A. In constructing questionnaires that use rating scales, one must be concerned with overcoming certain **rating errors** (also called **response biases** or **rater biases**).

B. **Halo Effect**: An observer forms a favorable impression of a person based on one central trait and extends that impression to all of the person's characteristics. Forced-choice scales were created to overcome this type of response bias.

C. **Leniency Bias**: Respondents rate someone who is very familiar, or someone with whom they are ego-involved, in an unrealistically positive manner. One way to overcome this error if using a graphic scale is to give only one unfavorable cue word; the rest of the range is then made up of favorable responses in different degrees.

D. **Central Tendency Bias**: The respondent is hesitant to give extreme ratings and instead clusters his or her responses around the center choice. This potential bias can be addressed by expanding the range of the scale.

E. **Ceiling or Floor Effects**: These occur when an extreme value is chosen on a pre-measure so that further change in that direction is restricted on the post-measure. Extending the scale can help to overcome this.

F. **Logical Error in Rating**: Respondents give similar ratings for variables or traits that they connect as logically related in their own minds but that may not occur together in the person being rated. The standard way to overcome this error is to construct very precise definitions and to make the instructions as explicit as possible.

G. **Acquiescent Response Set**: The tendency for some respondents (called **yea-sayers**) to go along with almost any statement. This bias is addressed by using both anti and pro items.

VIII. What Is the Semantic Differential Method?

A. The **semantic differential method** was developed as a way to study attitudes about the subjective (or representational) meaning of things in everyday life.

B. Respondents evaluate something using segmented graphic scales with bipolar cue words that represent the three primary dimensions of subjective meaning: **evaluation**, **potency**, and **activity**.

C. One or more lesser dimensions may be useful as well; these are **stability, tautness, novelty,** and **receptivity.**

D. Increasing the number of items generally increases the internal-consistency reliability of the instrument as a whole.

IX. What Are Likert Scales and Thurstone Scales?

A. The **summated ratings method**, developed by Rensis Likert, gives a one-dimensional picture of people's attitudes on controversial issues.

B. Attitude questionnaires created using the summated ratings method are known as **Likert scales**.

C. Steps in using the summated ratings method to construct a questionnaire:

1. Write a large number of statements on the controversial issue.
2. Have a sample of people from the target population indicate their evaluations of each statement, usually by means of a five-point numerical scale.
3. Identify the 20 or so statements whose responses best correlate with the total score (i.e., the sum of the scores for all of the items). These statements are then selected for the final questionnaire.

D. The **method of equal-appearing intervals**, developed by L.L. Thurstone, involves judges sorting statements into different piles that are assumed to be psychologically equidistant.

E. Attitude questionnaires developed using the method of equal-appearing intervals are also known as **Thurstone scales**.

F. Steps in using the method of equal-appearing intervals to construct a questionnaire:

1. A large number of statements are printed individually on slips of paper or index cards.
2. Judges then sort the statements into 11 piles numbered from 1 (labeled "most unfavorable statements") to 11 ("most favorable statements").
3. Scale values are obtained for each statement. They are usually calculated as the median of the responses of all the judges to that item.
4. In selecting statements for the final questionnaire, the idea is to choose those:
    a. That are most consistently rated by the judges.
    b. That are spread relatively evenly along the entire attitude range.

G. The attitude score for each respondent is the median scale value of the statements endorsed by the respondent.

X. How Are Items Prepared for a Questionnaire or an Interview?

A. Pilot testing one's questionnaire or interview is absolutely essential and can be used to identify potential questionnaire problems.

1. Are the items worded properly?
2. Does the way in which the items are worded and presented lead the respondent into giving an unrealistically narrow answer?
3. Are there **leading questions**, which can constrain responses and produce biased answers?

B. Identified problems can often be resolved with rewording or with a set of probing items instead of a single item.

C. Pilot testing can help answer the question of whether to use open-ended or more structured items (or a combination of both).

1. Exploratory questions about the respondent's experience of the questionnaire can be considered.
2. Respondents might also be asked about how strongly they believe their responses or how confident they are about them.

D. The **critical incident technique** can be used to prevent rambling in open-ended questions.

XI .How Are Face-to-Face and Telephone Interviews Done?

A. Questionnaires are convenient because:
    1. They can be administered to large numbers of people.
    2. They are relatively economical.
    3. They provide a type of "anonymity."

B. **Face-to-face interviews** have the following advantages:
    1. They provide an opportunity to establish a rapport with the participants and to stimulate the trust and cooperation needed to probe sensitive areas.
    2. They provide an opportunity to clarify questions if necessary.
    3. They allow flexibility in determining the wording and sequence of the questions, thereby giving the researcher greater control over the interview.

C. An **interview schedule** is a script of questions to be asked in the interview.

D. Steps in pilot testing the interview schedule:
    1. Identify the objectives of the interview.
    2. Formulate a recruitment strategy for locating potential interviewees.
    3. Structure the interview schedule.
        a. Write items and check each one for relevancy.
        b. Determine the ranges of responses for some fixed-choice items.
        c. Establish the best sequencing and wording of questions.
    4. Pilot-test the interview schedule and make appropriate revisions.

E. Sequence of questions should also be considered.

XII. Interviews by Telephone

A. Advantages include:
    1. Allows for a quick turnaround.
    2. Refusal rates are usually lower in telephone interviewing.

B. Disadvantages include:
    1. Many people use only mobile phones, but area probability sampling is restricted to households that have a land-line telephone linked to a specific geographical location.
    2. Interviewing is restricted to households that own a telephone and, then, to those that answer the telephone.
    3. Fewer questions (and less probing questions) can be asked because of the difficulty in establishing a rapport as compared to a face-to-face interview as well as people are more impatient to conclude a telephone interview.

C. Generally speaking, the same procedures are followed regardless of whether telephone or face-to-face interviewing is used but there is less time to establish rapport on the phone.

XIII. How Are Behavioral Diaries Used in Research?

A. A problem with the use of self-report measures to obtain autobiographical information stems from the limitations associated with one's ability to recall such information.

B. One way to overcome the problems associated with the reliance on one's memory is to use a **behavioral diary**. Participants are asked to keep a diary of events as they occur.
C. The assumption that behavioral diaries provide more reliable data than questionnaires or interviews that elicit answers to autobiographical questions has been supported by research.
D. However, some researchers have argued that behavioral diaries may lead individuals to be overly attentive to certain events, and therefore, under-report the occurrence of other behaviors or events.

LECTURE IDEAS AND ACTIVITIES

1. To illustrate the differences between open-ended and closed or structured questions, ask your students the same question using both formats. For example, ask students to describe their impressions of their school in their own words and then ask them to rate the school on a seven-point scale (1 = awful; 7 = terrific). Have students discuss the advantages and disadvantages of using these two question formats when asked essentially the same question—how do they feel about their school?

2. A valuable resource for obtaining examples of different types of self-report measures is *Measures of Personality and Social Psychological Attitudes*, edited by Robinson, Shaver, and Wrightsman. The editors provide reliability and validity information associated with each scale which may serve as an opening discussion for the next chapter.

Robinson, J. P., Shaver, P. R., & Wrightsman, L. S. (Eds.) (1991). *Measures of personality and social psychological attitudes.* New York: Academic Press.

3. One of the issues discussed in this chapter is the concern that respondents may be evasive in their answers, particularly if the questions are sensitive in nature. Gale Miller (1999) has developed a very effective demonstration of how evaluation apprehension concerns can lead to response biases in self-reporting. There are two parts to this activity. Miller first has her students take out a clean sheet of paper and number it from one to five. She then has her students take an "honesty" pledge by raising their right hand and repeating the following statement: "I promise, promise, promise, as a(n) (insert school mascot name) to tell the truth no matter what." Miller tells the students that she will be asking them a series of true/false questions. They are to indicate with a "T" or "F" on the paper whether or not the statement is true for them. She also encourages the students to keep their answers concealed from other students. Miller proceeds to ask her students to respond to the following statements:

1. I am a female.
2. I live in an apartment.
3. I have been what I consider to be legally drunk in the past six months.
4. I have engaged in unprotected sexual intercourse.
5. I have masturbated in the last six weeks.

The students hand in their papers so that the results can then be tallied. Before the results are revealed, Miller tells her class that they are going to respond to another series of questions. However, this time they will simply raise their hands if the statement is true for them. Once again, she has her students first take the "honesty pledge" before answering the questions. Miller then asks students to respond with a show of hands to the very same questions that were asked the first time (i.e., I am a female, etc.). We then compare the number of people who indicated with a show of hands whether or not they have masturbated in the past six weeks with the number of respondents who had affirmatively answered the question when they were responding in private on paper. These results usually lead to a lively discussion of why this self-report bias occurred and how a researcher might try to minimize this bias, especially when asking about sensitive issues.

Miller, G. A. (1999, August) *Teaching demonstration: Self-report bias lecture introduction to surveys, interviews, and questionnaires.* Paper presented at the annual convention of the American Psychological Association, Boston, MA.

4. One strategy for introducing self-report methods in your class is to have students develop their own self-report measure. For example, have your classes develop questionnaires concerning students' attitudes towards Valentine's Day, cell phones, and personal hygiene practices. Require students to use open-ended as well as fixed-choice items. For the fixed-choice items, have the students use the three types of rating scales discussed in the text (i.e., numerical, forced-choice, and graphic). This exercise is a useful tool for discussing the advantages and disadvantages of using open-ended and fixed-choice items, as well as the different types of rating scales and how they are useful for addressing certain rating errors. Upon completion of the questionnaire, have students also pilot-test the instrument. This provides an opportunity to discuss the importance of pilot-testing when finalizing the format of the self-report measure. This exercise can also be used to introduce the concepts of reliability and validity, which are discussed in the next chapter of the text.

5. To introduce as well as "personalize" the idea of rating errors, have students discuss their approach to completing self-report measures. For example, some students may admit that they never use the end points of a scale because those ratings seem so "extreme." With this admission, begin a discussion of the central tendency bias and how one can try to minimize this bias. To illustrate the leniency bias, mention that students tend to rate their professors as "above average" on teacher evaluation forms. Then present the question, "If all of the professors are considered to be above average, who is average?" This realization leads into a discussion of the leniency bias and how it can be minimized. The other rating errors in the text can be introduced in a similar manner.

6. John Ellard and T. B. Rogers (1993) have developed a list of guidelines (or "rules") students should follow when constructing questionnaires involving close-ended questions. To help students remember these rules, Ellard and Rogers describe them as being "The Ten Commandments of Question Writing," as follows:

1. Thou shalt not create double-barreled items.
2. Thou shalt not use "no" and "not" or words beginning with "un."
3. Thou shalt match the vocabulary used in items to the vocabulary of those who will respond to them.
4. Thou shalt not use complex grammatical forms.
5. Thou shalt have 40 to 60% true- or agree-keyed items.
6. Thou shalt not use redundant or irrelevant items.
7. Thou shalt not permit any loaded questions to appear in your questionnaire.
8. Thou shalt not mix response formats within a set of questions.
9. Thou shalt not permit a non-committal response.
10. Thou shalt pretest questions before collecting data.

Following Ellard and Rogers's suggestion, create a handout that presents these guidelines in a form suggestive of the biblical Ten Commandments. Ellard and Rogers have found that presenting these rules in this manner appears to enhance students' interest in and retention of these rules when constructing a questionnaire.

Ellard, J. H., & Rogers, T. B. (1993). Teaching questionnaire construction effectively: The ten commandments of question writing. *Contemporary Social Psychology, 17,* 17–20.

7. To illustrate the importance of pilot testing a newly developed instrument, discuss the case of the flawed Holocaust survey. It was widely reported in 1992 that a survey by the Roper Starch Worldwide organization for the American Jewish Committee showed that 22% of Americans had doubts concerning the veracity of the Holocaust while another 12% were not sure (Morin, 1994). Media reports of these findings created a serious concern in the public about what was being taught in the schools. However, a follow-up survey by the Gallup Organization found that only three to five percent of Americans doubted the occurrence of the Holocaust while 10% were not sure.

Why were there such discrepant findings between the two surveys? The answer appears that the differing results were due to the wording of the survey questions. For the Roper Starch survey, an important question read, "Does it seem possible or does it seem impossible to you the Nazi extermination of the Jews never happened?" There were at least two problems with the wording of this question (Morin, 1994). First, the question involves a double negative. Second, the word "extermination" may have been too strong in that individuals may have construed it to mean that all of the Jews were killed. The Gallup Organization conducted an experiment to see if the wording of the question affected the results. One group was provided with the original wording of the question. A second group was given this simplified version of the question, "Do you doubt that the Holocaust actually happened or not?" The Gallup Organization replicated the original survey's findings with the first question, but found that only 9% doubted its occurrence when presented with the second wording of the question.

Morin, R. (1994, March 20). That misleading Holocaust survey. *The Washington Post,* p. C3.

8. To demonstrate the utility of behavioral diaries in improving the accuracy of autobiographical data, have students estimate the number of hours they spend studying during a typical week. Then have the students keep track of the number of hours they actually spend studying during the following week. At the end of the week, have students compare the number of hours they reported that they study during a typical week and the number of hours they recorded as actually studying. Discuss the implications that discrepancies between these two reports might have with respect to the accuracy and reliability of data based on the respondents' recall abilities.

9. Arguing the importance of teaching survey methods along with experimental methods, Jan Yoder (1996) describes an activity one can use to help students experientially learn interviewing skills. Yoder first teaches her class how to conduct an interview by showing the film, *Interviewing* (Video Works, 1974). Yoder then provides students with a copy of an interview instrument. This provides Yoder with an opportunity to discuss issues such as differences between closed and open-ended questions. After discussing the instrument, Yoder divides the class into three-person groups to practice the interview schedule. Each member of the group takes turns at being the interviewer, the interviewee, and an observer who provides constructive feedback to the interviewer. Yoder found that this activity was beneficial for students to both learn and experience the skills necessary to conduct a successful interview.

Yoder, J. (1996). Teaching students to do interviewing. *Handbook of demonstrations and activities in the teaching of psychology* (Vol. 1, pp. 188–189). Washington, DC: American Psychological Association.

MULTIPLE-CHOICE QUESTIONS

1. Observations that ask research participants to look within themselves and describe their impressions are referred to as
   a. systematic observations.
   * b. self-report measures.
   c. secondary observations.
   d. exploratory measures. (82)

2. Self-report measures that are developed and are administered and scored according to certain rules are referred to as _____ measures.
   a. fixed-choice
   b. open-ended
   c. psychological
   * d. standardized (82)

3. Whenever Jack is asked to answer questions about himself, he is worried that his answers will portray him in an unfavorable light. As a result, Jack tends to answer questions in a way that he believes will give others a more positive impression of himself. Jack is experiencing _____ whenever he is asked to complete a self-report measure.
   a. an ethical conflict
   b. limited recall
   * c. evaluation apprehension
   d. a leniency bias (83)

4. Self-report measures that allow respondents to express their feelings and impressions in their own words are referred to as _____ measures.
   * a. open-ended
   b. fixed-choice
   c. standardized
   d. closed (85)

5. Renee asks students to describe their first experiences at college in their own words in order to gain an insight into the freshman year experience. Renee is using a(n) _____ measure.
   a. closed
   b. fixed-choice
   * c. open-ended
   d. autobiographical (85)

6. Self-report measures that provide respondents with fixed response options are referred to as _____ measures.
   a. open-ended
   b. precise
   * c. fixed-choice
   d. forced-choice (86)

7. Julie asks students to rate their freshman experience on a nine-point scale (1 = extremely unpleasant; 9 = extremely pleasant). Julie is using a(n) _____ self-report measure.
   a. limited
   b. open-ended
   c. forced-choice
   * d. fixed-choice (86)

8. _____ measures allow one the opportunity to freely express one's feelings whereas _____ measures do not.
   * a. Open-ended; fixed-choice
   b. Fixed-choice; open-ended
   c. Structured; fixed-choice
   d. Open-ended; unstructured (85-86)

9. The primary difference between attitudes and opinions is that
   a. attitudes are a guiding label and opinions are not.
   b. attitudes can be used with a Likert scale but opinions cannot.
   * c. opinions are verbal entities that can be measured directly, while attitudes are inferred entities.
   d. opinions influence participants' behavior in some way, while attitudes are designed to bring out explicit verbal responses. (88)

10. Anthony G. Greenwald and Mahzarin R. Banaji (1995) developed an approach intended to measure what they called implicit attitudes, using what they called the
   * a. Implicit Attitude Test (IAT).
   b. Minnesota Multiphasic Personality Inventory (MMPI).
   c. Forced-choice scale.
   d. Central tendency bias. (88-89)

11. Which of the following is an example of a projective test?
   a. Minnesota Multiphasic Personality Inventory
   * b. Thematic Apperception Test
   c. Thurstone attitude scale
   d. Likert scale (87)

12. Paul is asked to describe what he sees in a series of inkblots. Paul is completing a
   a. Thematic Apperception Test.
   b. Minnesota Multiphasic Personality Inventory.
   * c. Rorschach test.
   d. Likert scale. (87)

13. Andrew is given a series of pictures and asked to make up a story explaining each picture. Which personality measure is Andrew completing?
   a. MMPI
   * b. TAT
   c. Rorschach
   d. Thurstone (87)

14. An example of a structured personality measure is the
   a. Rorschach test.
   * b. Minnesota Multiphasic Personality Inventory.
   c. Thematic Apperception Test.
   d. Semantic Differential Scale. (101)

15. A type of rating scale that requires respondents to work with a sequence of defined numbers is the
   a. graphic scale.
   b. forced-choice scale.
   c. segmented scale.
   * d. numerical scale. (89)

16. Peter responds to an item on a measure by circling "5" which indicates that he "strongly agrees" with the item. Peter is most likely completing a measure using a _____ rating scale.
    a. graphic
  * b. numerical
    c. segmented
    d. forced choice (89)

17. Rachel develops an instrument in which respondents are supposed to indicate whether they strongly disagree, disagree, agree, or strongly agree to each question. Which type of rating scale is Rachel employing?
  * a. numerical
    b. forced choice
    c. graphic
    d. bipolar (89-90)

18. Rating scales that require respondents to choose between equally desirable or equally undesirable response alternatives are called
    a. graphic scales.
    b. numerical scales.
    c. Likert scales.
  * d. forced-choice scales. (90)

19. True-false questions utilize which type of rating scale?
    a. numerical
  * b. forced-choice
    c. graphic
    d. Likert (90)

20. Which type of rating scale may be effective in minimizing the halo effect?
  * a. forced-choice
    b. bipolar
    c. numerical
    d. graphic (90)

21. Barbara asks respondents to make a check mark on a straight line to indicate how they would rate themselves on several dimensions. Barbara then uses a ruler to transform these marks into numbers. Barbara is most likely using which type of rating scale?
    a. forced-choice
    b. Thurstone
    c. numerical
  * d. graphic (91)

22. Which rating error refers to a person forming a favorable impression based on one central trait and then extending this impression to all of the person's characteristics?
    a. leniency bias
    b. positivity bias
  * c. halo effect
    d. logical error in rating (90)

23. Because Janet is perceived by Hal as being warm, Hal also judges Janet to be friendly, kind, and sincere. This is an example of which rating error?
  * a. the halo effect
    b. the leniency bias
    c. the logical error in rating
    d. the central tendency bias (90)

24. Megan consistently tends to give all of her professors extremely positive teaching evaluations. Which rating error would best describe Megan's responses?
    a. halo effect
  * b. leniency bias
    c. central tendency bias
    d. acquiescent response set (92)

25. Barbara refuses to use the "not at all" and the "extremely" response options on a rating scale. Which rating error does this exemplify?
    a. acquiescent response set
    b. logical error in rating
    c. leniency bias
  * d. central tendency bias (92)

26. While making interpersonal judgments, Fred gives a person similar ratings on traits that he sees as being interrelated without serious consideration of whether the person actually exhibits all of these traits. This is an example of which rating error?
    a. halo effect
  * b. logical error in rating
    c. central tendency bias
    d. acquiescent response set (92)

27. Upon examination of a respondent's answers, Tom notices that the individual circled the "strongly agree" option for every item on the questionnaire. This is most likely an example of which response bias?
    a. logical error in rating
    b. leniency bias
    c. halo effect
  * d. acquiescent response set (92)

28. "Yea sayers" are respondents who exhibit which response bias?
    a. central tendency bias
  * b. acquiescent response set
    c. logical error in rating
    d. leniency bias (92)

29. Mark is asked to evaluate how he perceives certain career options along three different dimensions. Mark is most likely completing a
    a. Likert scale.
    b. Thurstone scale.
    c. numerical scale.
  * d. semantic differential. (93)

30. Which of the following is NOT one of the three primary dimensions measured using the semantic differential method?
  * a. importance
    b. evaluation
    c. potency
    d. activity (93)

31. Which standardized measure is developed using the summated ratings method?
    a. Thurstone scale
    b. Rorschach Test
  * c. Likert scale
    d. Thematic Apperception Test (95)

32. Which of the following scales is constructed using the method of equal-appearing intervals?
    a. Likert
    b. semantic differential
  * c. Thurstone
    d. TAT (95)

33. Carol is given a list of attitude statements and is asked to check the statements with which she agrees. Carol's attitudinal position is then determined by calculating the median of the scale values from the statements she endorsed. Carol is most likely completing which standardized measure?
    a. Likert scale
  * b. Thurstone scale
    c. TAT
    d. MMPI (95)

34. Which of the following self-report items would be considered to be a leading question?
    a. How satisfied are you with the President's economic policies to reduce unemployment?
    b. To what extent do you agree with the President's economic initiatives?
    c. How confident are you that the President's economic policies will be effective?
  * d. To what extent do you believe the President should be blamed for the high unemployment rate? (98-99)

35. In order to identify potential problems with one's questionnaire, one should always
    a. use as many respondents as possible.
  * b. pilot test the questionnaire.
    c. carefully proofread the questionnaire before administering it to others.
    d. pretend to be an actual respondent and personally complete the questionnaire. (98-99)

36. Interested in the freshman year experience, David asks freshmen to describe in their own words the first college class that they had attended. David is most likely using which self-report method?
    a. the projective test
  * b. the critical incident technique
    c. the semantic differential
    d. the structured interview (99)

37. Which of the following is NOT an advantage of face-to-face interviews?
    a. Face-to-face interviews provide an opportunity to establish rapport with the respondents.
    b. Face-to-face interviews provide an opportunity to help respondents with interpretation of the questions.
  * c. Face-to-face interviews provide an opportunity to challenge the respondents on questionable answers.
    d. Face-to-face interviews allow flexibility in determining the wording and sequence of questions. (99)

38. The script an interviewer uses to ask questions during a face-to-face interview is called a(n)
    a. self-report measure.
  * b. interview schedule.
    c. interview questionnaire.
    d. interviewer report. (100)

39. The procedures for developing a telephone interview schedule are _____ the procedures for developing face-to-face interview schedules.
    a. more complex than
    b. exactly opposite of
  * c. similar to
    d. unrelated to (101)

40. One way to overcome problems associated with the reliance on one's memory in self-reports is to use a(n)
    a. Thurstone scale.
    b. projective test.

* c. behavioral diary.
d. face-to-face interview. (101)

41. To study college students' eating behaviors, Diane asks students to keep a log of what they eat for one week. Diane is using which method of self-report?
a. tally sheets
b. Likert scale
c. semantic differential
* d. behavioral diaries (101)

SHORT ESSAY QUESTIONS

1. Why must researchers sometimes use self-report measures rather than exclusively relying on systematic observations of behavior?

2. What is the difference between open-ended and fixed-choice self-report measures? What are the advantages and disadvantages of each type of measure?

3. How are attitudes measured? What is the difference between an attitude and an opinion? Provide one example of a test that measures implicit attitudes.

4. Describe the three common types of rating scales.

5. What is a rating error? Describe three types of rating errors and how they may be controlled in a self-report measure.

6. What is the semantic differential method for studying attitudes?

7. Describe how a scale is created using the summated ratings method.

8. Describe how a scale is created using the method of equal-appearing intervals.

9. Why is pilot-testing important in the development of questionnaires? What problems might pilot-testing identify?

10. What advantages do questionnaires have over face-to-face interviews? What advantages do face-to-face interviews have over questionnaires?

11. What is an interview schedule? Describe the four steps in pilot-testing an interview schedule.

12. Why might a researcher ask subjects to keep diaries of experienced events rather than using questionnaires or interviews to investigate those events?

CHAPTER 6: *RELIABILITY AND VALIDITY IN MEASUREMENT AND RESEARCH*

CHAPTER OUTLINE

I. What Is the Difference Between Validity and Reliability?
- A. **Validity** refers to how well the measure or research design does what it purports to do.
  1. Considerations of the validity of a test include criterion validity, construct validity, and content validity.
  2. Internal and external validity are important considerations when evaluating research designs.
- B. **Reliability** refers to the extent to which observations or measures are consistent or stable.

II. What Are Random and Systematic Errors?
- A. There are two types of error:
  1. **Random error** refers to chance fluctuations, or haphazard errors.
  2. **Systematic error** (or bias) does not cancel out; rather it is slanted in a particular direction.
- B. In classical test theory, obtained scores are comprised of the theoretically "true scores" and random errors (or errors of measurement). The greater the random fluctuations, the less consistent or dependable (i.e., the less reliable) are the raw scores.
- C. The difference between systematic and random errors is that:
  1. Random errors are likely to cancel one another, on the average, over many repeated measurements.
  2. Systematic errors do not cancel one another, but instead affect all measurements in roughly the same way.

III. What Is the Purpose of Retest and Alternative-Form Reliability?
- A. **Test-retest reliability** is an estimate of the degree of fluctuation of an instrument, or of the characteristic it is designed to measure, from one administration to another.
  1. The test-retest reliability can be represented by a **correlation coefficient** between the scores on the test administered at two different times.
  2. The higher the test-retest coefficient, the more dependable or temporally stable is the instrument.
- B. A common concern is that the test-retest coefficient may be artificially inflated because of the respondents' familiarity with the test items.
  1. One way to avoid this problem is to create two statistically and theoretically comparable forms of the test with different items that are measuring the same content.
  2. The reliability of the set of scores, known as **alternate-form reliability**, is also assessed using the correlation coefficient.
- C. Test-retest reliability can be understood as a measure of stability whereas alternate-form reliability can be conceptualized as a measure of equivalence.

IV. What Is Internal-Consistency Reliability, and How Is It Increased?

A. **Internal-consistency reliability** indicates the degree of relatedness of the individual items. It is also called the **reliability of components**.

B. There are several ways of estimating internal-consistency reliability.

1. One traditional approach is to use the Spearman-Brown formula, which is based on the average intercorrelations of all the items.
2. Two other traditional approaches are **K-R 20** and **Cronbach's alpha coefficient**.

C. **Item-to-item reliability** is the estimate of the reliability of any single item on average.

D. To estimate the internal-consistency reliability of a multiple-item test, one uses the **Spearman-Brown Formula**. This formula is expressed by:

$$R^{SB} = \frac{nr_{ii}}{1 + (n-1)r_{ii}}$$

where n = the number of items in the test, and r_{ii} = the average intercorrelations of the items.

E. R^{SB} can be used to determine how the internal-consistency reliability will change by changing the number of items in a test.

V. What Are Acceptable Test-Retest and Internal-Consistency Reliabilities?

A. There is no simple answer to the question of what is an acceptable range of reliability.

B. The acceptable range depends on the situation in which the instrument is to be used and the objective of the research.

1. r_{ii} of the WAIS is .87; r_{ii} of the MMPI is .84; and for the Rorschach, r_{ii} is .86.
2. Retest r for the WAIS is .82; .74 for the MMPI; and .85 for the Rorschach.

VI. How Is the Reliability of Judges Measured?

A. Reliability is also a basic consideration in observational studies that use raters. B. **Judge-to-judge reliability** is the reliability of any single judge on average. It is the mean of the correlations between all pairs of judges.

C. One would use the Spearman-Brown formula to determine the reliability of the group of judges as a whole. This formula is now expressed as:

$$R^{SB} = \frac{nr_{jj}}{1 + (n-1)r_{jj}}$$

where n = the number of judges, and r_{jj} = the average judge-to-judge reliability.

VII. How Is Reliability Related to Replication and External Validity?

A. **External validity**, according to Shadish, Cook, and Campbell (2002, p. 82), refers to "inferences about the extent to which a causal relationship holds across variations in persons, settings, treatments, and outcomes."

B. Just as one is interested in the dependability of measurements, one is also interested in the dependability of causal generalizations in experimental research (or external validity) based on replicable findings.

C. A convenient way of evaluating whether a researcher has been able to replicate a set of research results, given the assumption that this replication will be relative rather than exact, is through a comparison of the effect sizes associated with each set of research results.

1. The **effect size** refers to the magnitude of the relation between the independent and dependent variables.
2. Research results are considered to be replicated when the effect size of the results from the replication study is similar to the effect size of the results of the original study.
3. Effect sizes can be compared statistically to rule out chance variation.

D. Broadly speaking, there are two categories of threats to the external validity of causal inferences:
1. Variables that were *not* in the experiment (i.e., variations in persons, settings, and treatments).
2. Variables that *were* in the experiment (e.g., operationalizing the variable of interest too narrowly, or using a specialized group of participants, or conducting the research in a setting that is quite unlike the circumstances to which one wants to generalize).

VIII. How Are Content and Criterion Validity Defined?

A. Does the test or measuring instrument actually do what it purports to do?
B. This assessment involves accumulating evidence in three categories, called content validity, criterion validity, and construct validity.
C. Content Validity
1. **Content validity** means that the test or questionnaire items represent the kinds of material (or content areas) they are supposed to represent.
2. Saying that a test or questionnaire has "good content validity" means that it adequately covers all major aspects of the content areas that are relevant.

D. Criterion Validity
1. **Criterion validity** refers to the degree to which the test or questionnaire is correlated with one or more outcome criteria.
2. In assessing criterion validity, researchers select the most sensitive and meaningful criterion in the present (called **concurrent validity**) or future (called **predictive validity**) and then correlate performance on the test or questionnaire with that criterion.

IX. How Is Construct Validity Assessed in Test Development?

A. The ability to discriminate is a vital characteristic of **construct validity**. It is generally considered the most "fundamental and all-inclusive validity concept, insofar as it specifies what the test measures" (Anastasia & Urbina, 1997, p. 114).
1. Campbell and Fiske (1959) suggest that researchers use two essential kinds of validity evidence to establish the construct validity of a test.
a. Test for convergence across different measures or manipulations of the same behavior (called **convergent validity**).
b. Test for distinctiveness between measures or manipulations of related but conceptually different traits or behaviors (**discriminant validity**).

B. Detailed Example: Crowne and Marlowe's Research

X. How Is Construct Validity Relevant to Experimental Design?

A. In research in which causal generalizations are the primary objective, **construct validity** refers to the validity of the hypothetical idea linking the independent and dependent variables. It also refers to the conceptualization of the independent and dependent variables.
    1. Among the more common threats to construct validity are vagueness in defining and operationalizing the concepts or variables of interest.
    2. It is a logical impossibility to do research without using constructs.
        a. Researchers need constructs to connect the operations they use in their studies to pertinent theory and to the way that causal generalizations will be used in practice.
        b. Constructs shape our perceptions and, because they also invariably have rich connotations, invite discourse and debate which stimulates further ideas for operationalizing and measuring these constructs.
        c. The "creation and defense of basic constructs" is the essence of what science is about.

XI. What Is the Importance of Statistical-Conclusion Validity and Internal Validity?

A. **Statistical conclusion validity** refers to whether certain statistical conclusions are well grounded, such as conclusions about the size of the effect or conclusions about statistical significance.

B. The essential characteristic of **internal validity** is the ability to rule out **plausible rival hypotheses**.
    1. Does an observed covariation between *X* and *Y* actually reflect a causal relationship from *X* to *Y*?
    2. There are a number of threats to internal validity which will be discussed in the next chapter.

LECTURE IDEAS AND ACTIVITIES

1. To introduce the notion of systematic error or bias, ask students whether they purposely set their clocks ahead of the actual time. Have students discuss how this knowledge of the clocks not being accurate affects their determination of the time. Relate this information to how scientists must also identify and correct for systematic biases that may consistently overestimate or underestimate the true value the scientist is trying to measure.

2. Michael Moore (1988) has developed a way of introducing the concept of measurement error. Moore provides students a sheet of paper with 50 lines of 25 different lengths. Students are asked to measure the length of each line. Moore then compiles the measurement data from each student and compares the mean student measurement against the "true" length of the line (according to his measurement). This activity leads to a discussion of why the measurements of each line were not exactly the same (they were, after all, measuring the same line) and the role that random and systematic errors have in measurement. Because each student is required to measure the same lines, you may also want to discuss the concept of interjudge reliability.

Moore, M. (1988). An empirical investigation and a classroom demonstration of reliability concepts. In M. E. Ware and C. L. Brewer (Eds.), *Handbook for teaching statistics and research methods* (pp. 41–42). Hillsdale, NJ: Lawrence Erlbaum.

3. To have students appreciate the importance of reliability when evaluating open-ended questions such as essay questions, J. Ronald Gentile (2000) has developed what he calls an "exercise in unreliability." Gentile divides the class into groups of three to five students. A volunteer is selected from each group. These volunteers are asked to go to a different room for a "special assignment." They are asked to take a piece of paper and pen with them. Once the volunteers have left the room, Gentile informs those remaining that these volunteers will be taking a surprise essay test consisting of one question relevant to the course content. The students remaining in the class are told that they will serve as an "examining committee" who will evaluate the volunteer from their group's answer to this question, as well as provide constructive feedback. After being given the question, each group is instructed to create a scoring system or rubric for evaluating their volunteer's answer.

While each group is developing their scoring system, Gentile leaves the classroom to meet with the volunteers in the other room. Gentile describes to the volunteers what the other students are doing and then informs them that they will be assisting as confederates in this activity on reliability. Gentile provides each volunteer with a somewhat sensible but not entirely correct answer to the essay question. Gentile then asks all the volunteers to copy the same exact answer in their own handwriting.

Once the groups have completed their scoring system for grading the essay answer, the volunteers are instructed to enter the classroom and give their supposed answer to their group to be evaluated. The volunteers again leave the room and each group is given five to ten minutes to score the essay and provide feedback on the answer for the volunteer. Once the groups have finished evaluating the essays, the volunteers return to the class to learn their grade on the essay exam. The volunteers are then asked to share with the rest of the class not only the grade they received from their group but also how their answer was scored. Gentile records the scores each volunteer received as well as how this score was determined. Once all of the volunteers have shared their score, it is revealed that every group evaluated the same exact essay answer. Noting the variability in the scores the different groups gave to the same exact answer, Gentile begins a discussion of the reliability and validity problems associated with essay scoring. You can certainly expand this discussion to the importance of reliability and validity considerations when evaluating responses to open-ended questions.

Gentile, J. R. (2000). An exercise in unreliability. *Teaching of Psychology, 27,* 210–212.

4. One way to introduce the concepts of reliability and validity is to have students actually evaluate the reliability and validity of a measurement instrument. Harold Takooshian describes an activity that has his students not only learn how to administer and score a standardized instrument, but also how to evaluate the reliability and validity of the test. This activity spans two class periods. During the first class period, students are asked to complete the Feminism

Survey (Takooshian & Stuart, 1983). Embedded within this survey are also questions that measure authoritarianism. Students are then provided with ten more surveys for them to distribute to others outside of class. They are instructed to administer these surveys to five individuals they suspect to be feminists and to five individuals they consider to be antifeminists. Students are also given a scoring sheet and a reliability worksheet that they must complete by the next class meeting. Using this reliability worksheet, students are asked to construct a scatter plot of the 10 pairs of odd-even scores as well as the Pearson correlation coefficient between these pairs (see Chapter 11). Students are also asked to calculate the mean feminist scores for the suspected feminist and antifeminist respondents in order to evaluate the validity of the scale. If the test is a valid measure of feminism, suspected feminists should have scores higher than suspected antifeminists. Similarly, the overall feminism score should not be positively correlated with one's authoritarianism score.

During the next class meeting, each student shares his or her high and low means as well as the calculated reliability *r* results. Using this information, Takooshian has the class evaluate the pattern of the results and draw conclusions concerning the reliability and validity of the Feminism Survey. Takooshian also uses this opportunity to show students how to calculate the composite reliability of the test by applying the Spearman-Brown correction to their calculated *r*. As an addition to this activity, you might have your students complete the Feminism Survey not only at the beginning of the first class meeting, but also at the beginning of the second class meeting. This allows you to also evaluate the test-retest reliability of the instrument.

Takooshian, H. (1999). Checking a test's reliability and validity. In L. T. Benjamin, B. F. Nodine, R. M. Ernst and C. B. Broeker (Eds.), *Activities handbook for the teaching of psychology* (Vol. 4, pp. 88–93). Washington, DC: American Psychological Association.

Takooshian, H., & Stuart, C. R. (1983). Ethnicity and feminism among American women: Opposing social trends? *International Journal of Group Tensions, 13,* 100–105.

5. George Howard, Paul Dunay, and Michael Crovello (1990) have developed an activity for teaching about construct validity that involves evaluating the construct validity of the midterm exam. Howard et al. collect several measures of students' knowledge of research methodology. Students are asked to complete a multiple-choice midterm exam, to respond orally to three essay-type questions concerning the class material, and to rate the perceived understanding of research methodologies for not only themselves but also for their classmates on a 10-point scale. Finally, so that the class can test for divergent validity, students also complete a test of sports trivia. Howard et al. use all of these measures to evaluate the convergent and discriminant validity of the midterm exam in order to demonstrate construct validation. This exercise can be challenging as the authors describe it as a classroom exercise "not for the faint of heart." Nevertheless, this exercise provides a variety of ways for introducing many of the concepts discussed in this chapter.

Howard, G. S., Dunay, P. K., & Crovello, M. T. (1990). An exercise in construct validation. In V. P. Makosky, C. C. Sileo, L. G. Whittemore, C. P. Landry and M. L. Skutley (Eds.), *Activities handbook for the teaching of psychology* (Vol. 3, pp. 27–32). Washington, DC: American Psychological Association.

6. Phillip Zimbardo (1981) describes a simple experiment that illustrates the importance of evaluating the internal validity of an experiment before one can make assertions based on the results of the experiment. Zimbardo begins with the hypothesis that "Males react faster than females." This hypothesis is almost guaranteed to capture the students' interest. After defining reaction time as the time interval between stimulus presentation and a person's reaction, Zimbardo produces his "reaction-time device." A simple ruler can be just as effective as Zimbardo's "reaction-time device." Zimbardo asks a member of the "slower" sex to come to the front of the room and extend her hand about level with the reaction-time device, with the thumb and forefinger about two inches apart. Without warning, Zimbardo drops the reaction-time device between the subject's thumb and forefinger. The subject will most likely catch the device. Zimbardo records how far the device dropped before being caught and operationalizes this measurement as "reaction time."

For the "faster" sex, Zimbardo repeats the procedure, but this time asks the male subject to extend his preferred hand. In addition, Zimbardo explicitly tells the subject what he is supposed to do (i.e., catch the device as soon as it is dropped), as well as provides a warning to the subject when the device is about to be dropped. Finally, Zimbardo allows the "faster" subject to have two trials, recording only the faster trial. Upon completion, Zimbardo declares that the "obvious conclusion" is that the hypothesis has been supported.

At this point the women students will probably and justifiably begin protesting, claiming that the procedure was biased. Have the protesting students specifically list their complaints with the "experiment," as well as what could have been done to minimize these biases. Discuss the importance of evaluating a study with respect to internal validity and how scientists are always trying to scrutinize their laboratory experiments to prevent the introduction of these rival hypotheses into their research. Next, ask students to assume that the experiment did, indeed, have internal validity. Have the students then consider the external validity of the study, which involved only two participants, one male and one female, from the class.

Zimbardo, P. G. (1981). Finding meaning in the method. In L. T. Benjamin, Jr. and K. D. Lowman (Eds.), *Activities handbook for the teaching of psychology* (Vol. 1, pp. 24–26). Washington, DC: American Psychological Association.

MULTIPLE-CHOICE QUESTIONS

1. Which of the following refers to how well a measure or research design does what it purports to do?
   * a. validity
   b. bias
   c. reliability
   d. error (106)

2. One must question an exam's _____ if it seems to measure more of one's ability to guess accurately than what one has learned in the class.
   a. reliability
   b. accuracy
   * c. validity
   d. bias (106)

3. Which of the following refers to the consistency or stability of a measurement?
   * a. reliability
   b. bias
   c. validity
   d. fluctuation (107)

4. An instrument that gives essentially the same results each time it is used, regardless of who administers it, is said to be high in
   a. accuracy.
   b. face validity.
   c. predictive validity.
   * d. reliability. (107)

5. Dennis consistently sets his watch 10 minutes ahead of the actual time. This is an example of
   a. random error.
   * b. systematic error.
   c. reliability.
   d. validity. (108-109)

6. _____ errors tend to cancel out over the long run whereas _____ do not.
   * a. Random; systematic
   b. Systematic; random
   c. Bias; systematic
   d. Bias; random (108-109)

7. Which type of reliability is concerned with the temporal stability of an instrument?
   a. split-half
   * b. test-retest
   c. internal-consistency
   d. Spearman-Brown (110)

8. Caroline administers a scale to a group of students and then asks them to complete the scale again a few days later. Which type of reliability is Caroline trying to assess?
   * a. test-retest
   b. internal-consistency
   c. inter-judge
   d. convergent (110)

9. To minimize the potential that scores on an instrument may be artificially inflated after people take the same test a second time, researchers sometimes assess _____ reliability rather than test-retest reliability.
   a. internal-consistency
   b. split-half
   * c. alternative-form
   d. item-to-item (110-111)

10. The degree of relatedness of the individual items on a test is the focus of _____ reliability.
   a. test-retest
   b. judge-to-judge
   * c. internal-consistency
   d. alternative-form (111)

11. Margaret calculates the average inter-item correlations for her scale and then uses the Spearman-Brown formula for her final calculations. Margaret is calculating the _____ reliability of the instrument.
   a. alternative-form
   * b. internal-consistency
   c. test-retest
   d. item-to-item (111-112)

12. The Spearman-Brown formula is used to determine the _____ reliability of an instrument.
   a. item-to-item
   * b. internal-consistency
   c. test-retest
   d. alternative-form (112)

13. A researcher constructs a five-item measure of attitudes toward national health insurance. The average intercorrelations among the items is $r_{ii} = .40$. Using the Spearman-Brown formula, he finds that $R^{SB} = .77$. What is the reliability of the scale as a whole?
    * a. .77
    b. .50
    c. .40
    d. Cannot be determined without knowing the sample size. (112)

14. Which of the following is used to describe the judge-to-judge reliability?
    a. the Pearson *r* correlation between each judge
    * b. the mean of all the Pearson *r* correlations between all pairs of judges
    c. Cronbach's alpha coefficient
    d. effect size (114-115)

15. What is face validity?
    * a. whether the instrument seems on the surface (or "face") to be measuring something relevant.
    b. whether the instrument can generalize a causal relationship across persons and settings
    c. the degree to which a measure correlates with other outcome measures
    d. the "ability to discriminate" (118)

16. Which of the following refers to the ability to generalize a causal relationship across variations in persons, settings, treatments, and outcomes?
    a. internal validity
    b. statistical conclusion validity
    * c. external validity
    d. construct validity (116-117)

17. David is afraid that the results from his experiment may not be able to be replicated in real life situations. David is concerned that his experiment lacks which type of validity?
    * a. external
    b. predictive
    c. construct
    d. internal (116)

18. To evaluate whether a set of scientific observations has been replicated, one can compare the _____ of the different sets of observations.
    a. validity
    b. test-retest reliability
    * c. effect size
    d. internal consistency (117)

19. If a test consists of questions that reflect the kinds of material of interest to the researcher, then the test is said to have good _____ validity.
    a. criterion
    b. construct
    c. convergent
    * d. content (118-119)

20. Jim is frustrated with a class because the questions on the most recent exam did not reflect the material that was discussed in class or in the text. The source of his frustration was this exam's apparent lack of
    * a. content validity.
    b. construct validity.
    c. divergent validity.
    d. convergent validity. (118-119)

21. When students are upset because they perceive an exam as being "unfair," they are most likely concerned with the test's _____ validity.
    * a. content
    b. criterion
    c. construct
    d. predictive (118-119)

22. The degree to which a measure correlates with other outcome measures is called
    a. discriminant validity.
    b. construct validity.
    * c. criterion validity.
    d. content validity. (119)

23. Sam is developing an instrument that would identify those students who may be at risk for flunking out of school. After administering this test to a group of students, he compares the results of the test with their current classroom performance. Sam is most likely trying to evaluate the instrument's _____ validity
    a. discriminant
    * b. concurrent
    c. content
    d. predictive (119)

24. Using a future criterion to evaluate the criterion validity of an instrument is called
    * a. predictive validity.
    b. concurrent validity.
    c. discriminant validity.
    d. content validity. (119)

25. To assess the validity of a college admissions test, a researcher will correlate the test scores with the students' first semester grade-point average. Which type of validity is the researcher investigating?
    a. concurrent
    b. discriminant
    c. convergent
    * d. predictive (119)

26. Convergent and discriminant validity are aspects of which type of validity?
    a. content
    * b. construct
    c. concurrent
    d. criterion (119-120)

27. Evaluating whether an instrument correlates highly with similar measures as well as whether the instrument does not correlate well with dissimilar measures is to determine which type of validity?
    a. content
    b. criterion
    c. convergent
    * d. construct (120)

28. The "ability to discriminate" is a characteristic of which type of validity?
    a. criterion
    b. convergent
    * c. construct
    d. content (120)

29. Which type of validity refers to whether a measure correlates highly with different measures of the same trait or behavior?
    * a. convergent
    b. divergent
    c. concurrent
    d. content (120)

30. Which type of validity refers to whether a measure distinguishes between related but conceptually distinct behaviors or traits?
    a. predictive
    b. construct
    c. convergent
    * d. discriminant (120)

31. Construct validity of an instrument is established by examining an instrument's _____ and _____ validity.
a. concurrent; predictive
b. concurrent; divergent
c. convergent; predictive
* d. convergent; divergent (120)

32. When one's research purpose is to make causal generalizations, one must be concerned with the _____ validity of how the variables are conceptualized.
a. face
* b. construct
c. internal
d. content (122)

33. After analyzing his data, a researcher concludes that there was no relationship between the independent and dependent variables in his study. However, he is concerned with whether or not his conclusions may be erroneous. This researcher is concerned with which type of validity?
a. internal
* b. statistical-conclusion
c. convergent
d. construct (123)

34. The ability to rule out plausible rival hypotheses in an experiment refers to which type of validity?
* a. internal
b. construct
c. external
d. statistical conclusion (123)

35. After the completion of an experiment, a researcher realizes that there is another explanation for the results other than the original hypothesis. The researcher realizes that this experiment is deficient with respect to which type of validity?
a. construct
b. external
* c. internal
d. statistical conclusion (123)

SHORT ESSAY QUESTIONS

1. What is the difference between validity and reliability? Why are they both of concern to researchers?

2. What is the difference between random and systematic error? Which error represents more of a threat to the nature of the conclusions that can be drawn from a study? Why?

3. What is the purpose of test-retest reliability and alternate-form reliability? In what types of situations would one use each type of reliability?

4. What is internal-consistency reliability? Why would one use the Spearman-Brown formula to assess this type of reliability?

5. How can the Spearman-Brown formula be useful in determining how long an instrument should be or how many judges one should use in a study?

6. What is meant by external validity? How might one try to demonstrate the external validity of a study?

7. How can effect sizes be used to evaluate whether the results from a study have been replicated?

8. What is content validity? How might one try to ensure that an instrument has good content validity?

9. What is criterion validity? What role do concurrent and predictive validity have in determining criterion validity?

10. What is construct validity in test development? What type of evidence does one use in the construct validation of an instrument?

11. Why would experimenters be concerned with construct validity when they try to make causal generalizations?

12. What is statistical-conclusion validity?

13. What is internal validity? Why is it important in trying to identify causal relationships?

14. Which two types of validity are concerned with whether or not one's causal inferences are correct? Why?

15. Which two types of validity are concerned with the strength or weakness of one's causal inferences? Why?

CHAPTER 7: *RANDOMIZED EXPERIMENTS AND CAUSAL INFERENCE*

CHAPTER OUTLINE

I. What is the Purpose of Randomized Experiments?
- A. **Randomized experiments** are experiments in which allocation of sampling units (participants) to groups or conditions is done by a process of random assignment.
- B. Although randomized experiments are often referred to as "the gold standard" in causal inference, this procedure is not flawless.
- C. There are traditionally three reasons for using **random assignment** (also called **randomization**).
  1. It provides a safeguard against the possibility of experimenters' subconsciously letting their opinions and preferences influence which of the **sampling units** will receive any given treatment.
     a. **Sampling units** is a general way of referring to the participants, subjects, groups, or objects being studied.
     b. The term **treatment** is another name for the manipulated variable.
  2. Random assignment distributes the characteristics of the sampling units over the treatment and control conditions in a way that will not bias the outcome of the experiment.
  3. Random assignment permits the computation of statistics that require certain characteristics of the data.

II. How Is Random Assignment Accomplished?
- A. Experimenters use a variety of procedures to achieve random assignment.
  1. Statisticians speak of random assignment *rules* (or plans), for example:
     a. Assignment to treatment and control group may be by the flip of a coin to assign each participant.
     b. Assignment may be made using a table of random digits.

III. What are Between-Subjects Designs?
- A. **Between-subjects design** involves exposing subjects to one condition each.
  1. Another statistical name for the between-subjects design is **nested design**, because the subjects are "nested" within their own groups or conditions.
  2. A popular way of analyzing the data from a two-condition between-subjects design is to calculate a *t* test for independent samples (see Ch. 13).
  3. A typical procedure for independent dichotomous counts would be a chi square ($x^{2)}$ procedure (discussed in Chapter 15).
  4. Between-subjects designs are not limited to two groups or two conditions.

IV. What Is the Formative Logic of Experimental Control?
- A. Mill's Methods and the Logic of Experimental Control
  1. **Mill's methods** is the name given to certain "logical methods" popularized by the nineteenth-century English philosopher John Stuart Mill.

B. Two of Mill's methods together form the logical basis of causal inferences made in all two-group between-subjects randomized experiments.
    1. The **method of agreement** states, "If *X*, then *Y*." *X* symbolizes the presumed cause and *Y* symbolizes the presumed effect.
        a. In other words, *X* is a sufficient condition of *Y*, meaning that it is adequate to bring about the effect.
        b. The experimental group corresponds to Mill's method of agreement.
    2. The **method of difference** states, "If not-*X*, then not-*Y*."
        a. In other words, *X* is a necessary condition of *Y*.
        b. The control group corresponds to Mill's method of difference.
        c. The control group can be a zero control group or a **placebo control group**.

V. What Are Within-Subjects Designs?

A. **Within-subjects design** involves exposing subjects to all of the experimental conditions.
    1. Within-subjects designs involving only two conditions call for a *t* test for nonindependent samples (see Ch. 13).
    2. Within-subjects designs are also called **repeated-measures designs** as well as **crossed designs**.
    3. Within-subjects designs are not limited to two groups.

VI. What Are Factorial Designs?

A. **Factorial designs** are research designs with at least two factors, or variables of interest, and one or more levels of each factor.
    1. It is important to keep in mind that it is not just the research design that determines the statistical procedure used to analyze the results. Statistical procedures should always be guided by one's hypotheses or questions of interest.
    2. Randomized designs can have more than two factors and more than two levels within each factor, although these designs may be stretching the number of subjects rather thinly.

VII. What Is Meant By Counterbalancing the Conditions?

A. A two-factor design in which one factor is between-subjects and the other is within-subjects is an example of a mixed factorial design.

B. A problem with within-subjects designs is that the order in which the treatments are administered may be confounded with the treatment effects. To minimize this problem, researchers would use **counterbalancing,** which means rotating the sequences.

C. A specific statistical design that has counterbalancing built in is called the **Latin square design**. It is characterized by a square array of letters or numbers (representing treatment conditions), where each letter appears once and only once in each row and in each column.

VIII. Why Is Causality Said To Be "Shrouded in Mystery"?

A. Aristotle suggested that questions of causation can be answered in four distinct ways:

1. The **material cause** concerns the substance or substances necessary for the movement or coming into being of the effect.
2. The **formal cause** refers to the plan or development that gives meaning to the event.
3. The **final cause** refers to the objective or end purpose of the event.
4. The **efficient cause** is the activating force or event that was responsible for the effect. It is this cause, in particular, that experimenting scientists frequently have in mind when they theorize that a manipulated treatment "produces an effect."

B. The rules of establishing efficient causality are based on three observations of the eighteenth-century Scottish philosopher David Hume. These are contiguity, priority, and a constant conjunction.

IX. Who Do Scientists Logically Puzzle Out Efficient Causality?

A. The rules of establishing efficient causality are based on Hume's observations and have been boiled down to three essential criteria:

1. There must be evidence that the independent variable and the dependent variable are mutually related (**covariation** is present).
2. There must be evidence of **temporal precedence** to support the assumption that the presumed cause actually came before the presumed effect.
3. Attempts are made to rule out plausible rival hypotheses that may undermine causal interpretations. That is, scientists look for evidence of **internal validity**.

B. Scientists find that they must settle for the most compelling evidence for determining causality, even if the evidence is inconclusive. Therefore, causal inference is always subject to some degree of uncertainty.

X. What Conditions Pose a Threat to Internal Validity?

A. Quasi-experimental research is research in which the designs resemble randomized experiments but do not use random assignment. This section covers six general categories of threats to internal validity that occur in quasi-experimental research: 1) biased selection; 2) bias due to history; 3. bias due to maturation; 4) bias due to attrition; 5) bias due to testing; and 6) instrumentation bias.

B. Preexperimental designs are research designs that lack the experimental control necessary to evaluate plausible threats to validity.

C. **Biased selection** refers to how the sampling units assigned to different conditions were selected for those conditions. The term biased implies that the selection procedure resulted in groups that, even before the experimental intervention or manipulation, were systematically dissimilar in respondent characteristics relevant to the observed outcome. In the **one-shot case study**, all subjects are exposed to the treatment and then observed or measured.

1. This design is symbolized as **X-O**, where X = exposure of a treatment group to an event or experimental variable, and O = an observation or measurement.

2. This design is deficient because it does not allow for a comparison with subjects who were not exposed to the treatment.

D. **Bias due to history** refers to the presence of an event other than (but typically concurrent with) the treatment, that may be responsible or partly responsible for the observed effect. In the **one-group pre-post design**, subjects are measured before and after exposure to the experimental treatment.
   1. The design is symbolized as **O-X-O**.
   2. This design is still considered preexperimental because of the lack of comparison conditions (i.e., beyond the pretreatment measure).

E. **Bias due to Maturation** refers to certain intrinsic changes in the research participants, such as their growing older, wiser, stronger, or more experienced between the premeasurement and the postmeasurement. It becomes a threat to internal validity when it is not the variable of interest but the inferred causal relationship is nevertheless confounded by the presence of these changes.

F. **Bias due to attrition** refers to the differential loss of units in some conditions, if the remaining units are dissimilar in the treatment and control groups.

G. **Bias due to Testing** implies that being measured, tested, or observed initially (or repeatedly) can affect subsequent performance on the dependent variable.

H. **Instrumentation bia**s refers to the possibility that the posttreatment effect that was measured or observed was due to changes in the measuring instrument.

XI. What Are Artifacts In Research?

A. **An artifact** is regarded as a type of error that occurs systematically rather than randomly and, if ignored or left uncontrolled or uncorrected, can jeopardize the validity of conclusions.

B. Artifacts can affect not only internal validity, but also construct and external validity,

C. Subject-related artifacts relate to the use of humans as research subjects.
   1. **Demand characteristics** refer to cues that are unknowingly or unwittingly communicated to the subject concerning what behaviors are expected from the subject in the experiment.
      a. The **good subject** is the kind of research participant who is sensitive to demand characteristics and tries to give experimenters what they seemingly want to find.
      b. **Evaluation apprehension** is a condition where human subjects in psychological investigations are apprehensive about being evaluated and therefore are motivated to "look good" rather than help the cause of science.
   2. In order to identify potential demand characteristics that may be inadvertently operating in an experiment, a researcher may employ the use of **quasi-control subjects**, "coinvestigators" who step out of their participant role to reflect on the context of the experiment.

D. Experimenter-related artifacts are sources of bias resulting from the uncontrolled intentions or actions of the experimenters themselves.

E. The **experimenter expectancy effect** reflects the acknowledgement that an experimenter's expectations can lead to self-fulfilling prophecies.

F. One strategy for minimizing the experimenter expectancy effect is to use **blind experimenters**. Ideally trials can use **double-blind procedures,** that is neither the human participants nor the experimenters know which individuals are in the experimental and control groups.

G. An **expectancy control design** is a simple factorial design that not only assesses whether an expectancy effect is present, but also allows a direct comparison of that effect with the phenomenon of theoretical interest.

LECTURE IDEAS AND ACTIVITIES

1. The focus of the previous chapters has been on methods of data collection. This chapter now turns to the actual designs of studies. Peter Fernald and Dodge Fernald (1990) have developed an exercise that can be useful for helping students to learn how to identify research designs that would be most appropriate for addressing different research problems. Fernald and Fernald provide their students with a list of ten statements (e.g., (1) Jogging increases lung capacity; (2) The soul remains after death; (3) Individuals having one or more significant hobbies report more job satisfaction than individuals having no hobbies; and (4) Unmarried cab drivers talk more with their customers than do married cab drivers). Students are to decide which research strategy would be most appropriate for each question: naturalistic observation, clinical, survey, experimental, or not empirically testable. This exercise can be easily adapted to suit your needs. For example, you may want to select statements that would be applicable to the different research designs discussed in the text. For those interested, the complete list of statements can be found in Fernald and Fernald's article.

Fernald, P. S., & Fernald, L. D. (1990). Selecting appropriate research methods. In V. P. Makosky, C. C. Sileo, L. G. Whittemore, C. P. Landry and M. L. Skutley (Eds.), *Activities handbook for the teaching of psychology* (Vol. 3, pp. 33–34). Washington, DC: American Psychological Association.

2. David Watson (1990) has developed a simple demonstration for illustrating the utility of random assignment for controlling potential threats to internal validity stemming from the assignment of subjects to experimental conditions. Watson instructs his class that he has developed a new way of coaching basketball that will lead to better teams. To demonstrate the efficacy of this new coaching method, he proposes that an experiment be conducted using two groups of subjects. One group will be exposed to the new coaching method, the other one will not. Both groups will then play each other with the expectation that the new coaching method group will do better. (NOTE: This exercise can also be used to introduce the basic logic of experimental designs.) Watson then discusses a variable that the experimenter has no control over: the height of the players. He describes how random assignment using a coin toss can be used to assign players of differential height to each team to eliminate height as a possible confounding variable in the experiment.

You may slightly modify this demonstration in your classes. After soliciting volunteer "subjects" from the class to "participate" in this experiment, assign the taller students to the new

coaching method group and the shorter students to the control group. Then declare that the experimental group will beat the control group in the tournament because the new coaching method is superior to the traditional coaching methods. At this point students begin to protest, pointing out that there is an obvious height difference between the groups. This leads into a lively discussion of how a bias can be introduced into an experiment through the assignment of subjects as well as how this bias may be eliminated. To further demonstrate the rationale behind the use of random assignment, randomly assign these volunteers to the treatment conditions and then compare the mean height of each group. After repeating this procedure several times, students begin to see the utility of random assignment in trying to eliminate a potential confounding variable.

Watson, D. L. (1990). A neat little demonstration of the benefits of random assignment of subjects in an experiment. In V. P. Makosky, C. C. Sileo, L. G. Whittemore, C. P. Landry and M. L. Skutley (Eds.), *Activities handbook for the teaching of psychology* (Vol. 3, pp. 3–4). Washington, DC: American Psychological Association.

3. David Stang (1981) describes another way that the random assignment of subjects can be beneficial for minimizing a potential confounding variable. Using what he describes as a "one-potato-two-potato" technique of haphazard assignment, Stang assigns students to one of six groups and then calculates the mean height for the students in each group. He then creates two groups from these six groups and then recalculates the mean height for these two new groups. Based on a comparison of the group means from these two parts of the exercise, Stang begins a discussion of the benefits of random assignment in creating similar groups as well as the notion that group similarity can be increased by randomly assigning more people to each group.

Stang, D. J. (1981). Randomization. In L. T. Benjamin, Jr. and K. D. Lowman (Eds.), *Activities handbook for the teaching of psychology* (Vol. 1, pp. 18–19). Washington, DC: American Psychological Association.

4. To introduce their students to the logic associated with experimental design, Hank Rothgerber and Eric Day (1999) have their students participate in a 2 × 2 factorial between-subjects study. This study uses jelly beans to examine the impact of expectations on one's perceptions. First, Rothgerber and Day bring to class two flavors of jelly beans. One flavor is a typical flavor (e.g., cherry) while the other flavor is atypical (e.g., buttered popcorn). The flavor of the jelly bean represents an independent variable or factor. Individual jelly beans are placed into envelopes. Half of the envelopes contain a note explicitly indicating flavor of the jelly bean contained in the envelope while the other half of the envelopes do not contain this information. The presence or absence of this information represents the second independent variable. After students are randomly assigned to one of the four groups, they are instructed to open the envelope and eat the jelly bean. They are then to rate on a five-point scale (1 = very little, 5 = very much) to what degree did the jelly bean meet their taste expectations. After collecting the students' ratings, Rothgerber and Day debrief the students and begin a discussion of the important concepts that were associated with the research design they just employed (e.g., the use of random assignment; the use of factorial designs to test for main effects along with interactions) as well as alternative

ways of explaining the obtained results. This activity is also useful for introducing the analysis of variance as a statistical method for analyzing data (see Chapter 14).

Rothgerber, H., & Day, E. A. (1999). Using jelly beans to teach some concepts in research methodology. In L. T. Benjamin, B. F. Nodine, R. M. Ernst and C. B. Broeker (Eds.), *Activities handbook for the teaching of psychology* (Vol. 4, pp. 69–73). Washington, DC: American Psychological Association.

5. To demonstrate that causation can have many different meanings, have students list as many different "causes" for them coming to class that day as they can. Students may initially begin listing causes such as wanting to graduate or because attendance is a course requirement. Suggest other possible causal factors such as the fact that the clock hands were pointing in a certain direction (e.g., clock indicated it was 10 a.m.). Once students have generated these lists of causes for their class attendance, categorize the causes according to Aristotle's four kinds of causation. Finally, discuss why scientists focus on efficient causation rather than the other three kinds of causation.

6. It may be beneficial to demonstrate experimental design logic by designing a simple experiment and have your students identify potential threats to the internal validity of the experiment. For example, Samuel Cameron, Jack Christiano, and Bernard Mausner (1981) suggest a simple experimental design involving heart rate that can be used to introduce the logic behind the experimental method, as well as how a simple measure of physiological arousal such as heart rate can be confounded by other factors such as thought processes and physical activity.

Students are assigned the role of either experimenter or subject. Armed with a stopwatch, the experimenter takes the subject's resting pulse rate and repeats this measure after one minute. The subject is then asked to read aloud from a handout, after which the subject's pulse rate is measured. After three minutes, the subject's resting heart rate is once again measured. Finally, the subject is asked to run in place for 30 seconds, after which the subject's pulse rate is measured for a final time. Mean pulse rates for each condition are calculated and compared. Cameron et al. then begin a discussion concerning what the independent and dependent variables were in this study as well as what other variables may have affected the dependent variable (pulse rate). This exercise can be useful for illustrating Campbell's different threats to internal validity. For example, have students discuss whether instrumentation may be a confounding factor if students became more proficient at taking a pulse rate during the course of the experiment. Maturation may also be a confounding factor if the subjects' anxiety levels have decreased during the course of the experiment.

Cameron, S., Christiano, J., & Mausner, B. (1981). Experimental design: Varying heart rate. In L. T. Benjamin, Jr. and K. D. Lowman (Eds.), *Activities handbook for the teaching of psychology* (Vol. 1, pp. 10–11). Washington, DC: American Psychological Association.

7. Paul Solomon (1988) has proposed another method for introducing the logic of experimental designs as well as demonstrating the relevance of the experimental method. Solomon has his students develop experiments to test advertising claims. For example, Solomon describes three class experiments that were developed to evaluate claims of toilet tissue softness, effectiveness of two antiperspirants, and preferences for different cola sodas. This exercise can easily be modified to fit the students' interests and to introduce the various research designs discussed in the text.

Solomon, P. R. (1988). Science and television commercials: Adding relevance to the research methodology course. In M. E. Ware & C. L. Brewer (Eds.), *Handbook for the teaching of statistics and research methods* (pp. 146–150). Hillsdale, NJ: Lawrence Erlbaum.

8. Carolyn Stierhem (1981) has developed a classroom demonstration of the experimenter expectancy effect. Three students are selected as "experimenters." The remaining students are assigned to one of three groups, with a minimum of ten subjects in each group. The task for the subjects in each group is to guess which number the experimenter is thinking, a 0 or 1, for ten trials. The experimenters are provided with a random sequence of 0s and 1s and are told to record whether the subject is correct in the guess. In addition, each experimenter is given one of three sets of instructions that they are told not to share with anyone. The first experimenter is told to concentrate hard on the number for each trial and to expect a high degree of accuracy, as some people have good ability to do well on these types of tasks. The second experimenter is told not to expect a high degree of accuracy, as any correct answers will be due to pure chance. Finally, the third experimenter is not told any additional information. This experimenter's group is intended to serve as the control group.

After each experimenter conducts the experiment in a quiet place, the mean number of correct guesses for each trial is calculated and compared with the other experimenters. Because the only differences between the groups were the experimenter's instructions, any differences between the groups can be attributed to an experimenter expectancy effect. Discuss with the students how the experimenter may have unconsciously communicated his or her expectations to the subjects and what problems this artifact may create in an experiment.

Stierhem, C. (1981). Experimenter expectancy. In L. T. Benjamin, Jr. and K. D. Lowman (Eds.), *Activities handbook for the teaching of psychology* (Vol. 1, pp. 20–21). Washington, DC: American Psychological Association.

MULTIPLE-CHOICE QUESTIONS

1. Randomized experiments are experimental designs that involve
   a. the manipulation of variables.
   b. the use of humans as research participants.
   * c. the use of random assignment to treatment conditions.
   d. the strict adherence to ethical guidelines. (129)

2. In order to avoid bias in how subjects are assigned to the different treatment conditions, experimenters should use
   * a. random assignment.
   b. counterbalancing.
   c. random sampling.
   d. self-selection procedures. (130)

3. Which of the following is NOT a reason given for incorporating random assignment into one's research design?
   a. Random assignment permits the use of certain statistical tests.
   b. Random assignment prevents possible biases in determining which treatment a subject receives.
   * c. Random assignment increases the reliability and validity of one's experimental design.
   d. Random assignment distributes the characteristics of the sampling units over all experimental conditions. (130)

4. Experimental designs that expose subjects to only one treatment condition each are referred to as
   a. randomized designs.
   b. repeated-measures designs.
   c. within-subject designs.
   * d. between-subjects designs. (131)

5. Tom believes that studying while listening to music will improve test performance. To test this hypothesis, Tom has one group of participants listen to music while studying for a spelling test. A second group of participants study for the test in silence. Tom then compares the spelling test performance for both groups. Tom's study is best described as which type of experimental design?
   * a. between-subjects
   b. within-subjects
   c. repeated measures
   d. simple factorial (131)

6. Between-subjects designs are also referred to as
   * a. nested designs.
   b. matched designs.
   c. counterbalanced designs.
   d. crossed designs. (132)

7. Between-subjects designs in which participants are assigned to treatment conditions by means other than randomization are referred to as

a. cross-lagged designs.
b. longitudinal designs.
* c. nonequivalent-groups designs.
d. single-case experimental designs. (131)

8. Experimental designs that expose subjects to more than one treatment condition each are referred to as
a. expectancy control designs.
* b. repeated-measures designs.
c. nested designs.
d. between-subjects designs. (134)

9. Because he is limited in the number of subjects he can recruit, Tom decides to test his hypothesis that studying to music improves test performance by comparing individuals' test performance after studying to music to their test performance after studying in silence. This is an example of which type of design?
a. simple factorial
b. between-subjects
* c. within-subjects
d. nested (134)

10. Within-subjects designs are also called
* a. crossed designs.
b. nested designs.
c. counterbalanced designs.
d. matched designs. (134)

11. The order in which the treatments are administered can unintentionally create problems for the experimenter who is employing which research design?
a. simple factorial designs
* b. repeated-measures designs
c. between-subjects designs
d. nested designs (136)

12. An experimental design that administers one or more levels of one independent variable in combination with two or more levels of another independent variable is called a
a. between-subjects design.
b. within-subjects design.
c. simple randomized design.
* d. simple factorial design. (135)

13. To minimize problems in repeated-measures designs that can result from systematic differences between successive measurements, a researcher should use a method called
    a. random assignment.
  * b. counterbalancing.
    c. nesting.
    d. expectancy control. (136)

14. Notions of causality based on the substance or substances necessary for the movement or coming into being of the effect are referred to as _____ causes.
    a. formal
  * b. material
    c. final
    d. efficient (137)

15. When asked to explain what caused there to be a new library on campus, Dave explains that the mortar and bricks are responsible. According to Aristotle, Dave is focusing on the _____ cause for why there is a new library on campus.
  * a. material
    b. formal
    c. final
    d. efficient (137)

16. An explanation of causality based on the plan or development that gives meaning to the event is an example of which type of causality?
  * a. formal
    b. efficient
    c. final
    d. material (137)

17. When asked what caused there to be a new library on campus, Dave explains that there was a desire to enhance the quality of the students' education. According to Aristotle, Dave is focusing on the _____ cause for why there is a new library on campus.
    a. formal
    b. efficient
    c. material
  * d. final (137)

18. Notions of causality that refer to the objective or end purpose of an event are referred to as _____ causes.
    a. efficient
    b. material
    c. formal
  * d. final (137)

19. Dave explains that what caused there to be a new library on campus was the implementation of the architectural plans. According to Aristotle, Dave is focusing on the _____ cause for why there is a new library on campus.
    a. material
  * b. formal

c. final
d. efficient (137)

20. The activating force or event that is responsible for an effect is the _____ cause of that event.
* a. efficient
b. final
c. formal
d. material (137)

21. Dave explains that what caused there to be a new library on campus was the large donation by a wealthy alumnus of the university. According to Aristotle, Dave is focusing on the _____ cause for why there is a new library on campus.
a. formal
b. final
* c. efficient
d. material (137)

22. All of the following are criteria for determining efficient causation EXCEPT
a. temporal precedence.
* b. external validity.
c. covariation.
d. internal validity. (138)

23. By establishing that the independent variable and dependent variable are correlated with each other, one has satisfied which criteria for establishing efficient causality?
a. internal validity
b. temporal precedence
c. random assignment
* d. covariation (138)

24. Which of the following evidence is particularly difficult to obtain when one is trying to determine efficient causality based on relational research?
a. internal validity
b. covariation
* c. temporal precedence
d. external validity (138-139)

25. Ruling out plausible rival hypotheses is central to satisfying which criteria for establishing efficient causality?
    * a. internal validity
    b. covariation
    c. temporal precedence
    d. external validity (139)

26. Mill's method of agreement implies that *X* is a(n) _____ condition of *Y*.
    a. necessary
    * b. sufficient
    c. important
    d. activating (133)

27. Mill's method of agreement corresponds to which group in a simple research design?
    a. the placebo group
    b. the control group
    * c. the experimental group
    d. the randomized group (133)

28. Mill's method of difference implies that *X* is a(n) _____ condition of *Y*.
    a. activating
    b. sufficient
    * c. necessary
    d. adequate (133)

29. Mill's method of difference provides the rationale for including a(n) _____ group in an experimental design.
    a. treatment
    b. experimental
    * c. control
    d. expectancy control (133)

30. In a drug study, one group of subjects receives a sugar pill rather than the pill containing the ingredients under investigation. This "sugar pill" group is referred to as the _____ control group.
    * a. placebo
    b. experimental
    c. zero
    d. expectancy (133-134)

31. Research designs that do not employ the use of randomization are referred to as
    a. nonrandomized experimental designs.
    b. correlational designs.
    * c. quasi-experimental designs.
    d. pseudo-experimental designs. (139)

32. Which of the following is NOT a characteristic of quasi-experimental designs?
    a. the use of treatment conditions
    * b. the use of randomization to allocate sampling units
    c. the use of outcome measures

d. the recruitment of sampling units (139)

33. Which of the following issues makes inferences of causation more risky in quasi-experimental designs?
a. lack of treatment manipulations
* b. nonrandom assignment to treatment conditions
c. unreliable outcome measurements
d. the use of both human as well as nonhuman sampling units (139)

34. To assess the effectiveness of a special lesson plan, a teacher implements the lesson plan and then evaluates the effectiveness of the plan by analyzing the students' test scores upon completion of the learning unit. This is an example of which type of research design according to Campbell and his colleagues?
a. simple factorial design
b. one-group pre-post study
* c. one-shot case study
d. posttest-only control-group design (140)

35. Which of the following is an example of a preexperimental design?
* a. the one-group pre-post study design
b. the Solomon design
c. the pre-post control-group design
d. the posttest-only control-group design (140)

36. An educator is interested in whether including a physically challenged student in a kindergarten class will enhance the other students' level of compassion. The educator assesses the students' level of compassion at the beginning of the year, before the physically challenged student was assigned to the classroom. At the end of the school year, the educator reassesses the students' level of compassion and notes whether overall compassion has increased or decreased. This is an example of which type of research design according to Campbell and his colleagues?
a. the posttest-only control group design
b. the one-shot case study design
c. an expectancy control design
* d. the one-group pre-post study (140)

37. The control group was measured using a different measuring tape than the treatment group. This is an example of which potential bias?
a. bias due to testing
* b. instrumentation bias
c. bias due to history
d. biased selection (141)

38. I selected my 7th grade classmates as the control group and a group of 7th graders from another school as the treatment group. What bias has a good chance of impacting this study?
    a. bias due to history
    b. the posttest-only control-group design
    c. bias due to maturation
    * d. biased selection (140)

39. While participating in a timed experiment, the participants in the experimental group had to leave the room because of a fire drill. Which potential threat to internal validity would this example illustrate according to Campbell and his colleagues?
    a. maturation
    b. selection
    c. instrumentation
    * d. history (140)

40. While conducting a year-long study on friendship patterns in the sixth grade, a researcher discovers that some of the students had reached puberty during that year while others had not. Which potential threat to internal validity would this example illustrate according to Campbell and his colleagues?
    a. selection
    b. instrumentation
    * c. maturation
    d. history (141)

41. A researcher becomes aware that the judges he is using to make observations are becoming more proficient at their task. This improvement among the judges might suggest that _____ represents a potential threat to the internal validity of the study.
    * a. instrumentation
    b. history
    c. maturation
    d. selection (141)

42. A researcher notices that the participants assigned to the experimental group were less likely to attend the second session in his experiment than those in the control group. The researcher is concerned that _____ might represent a potential threat to the internal validity of her study.
a. history
* b. attrition
c. instrumentation
d. maturation (141)

43. Which of the following refers to a finding that results from conditions other than those intended by the experimenter?
a. selection bias
* b. artifact
c. good subject effect
d. experimenter expectancy effect (141-142)

44. Unintended cues that can influence a subject's behavior in an experiment are referred to as
a. expectancy cues.
* b. demand characteristics.
c. residual effects.
d. quasi-experimental manipulations. (142)

45. The observation that sometimes subjects are extremely willing to complete a meaningless task in order to help the experimenter illustrates the idea of
a. the experimenter expectancy effect.
b. the subject selection bias.
* c. the good subject effect.
d. the Hawthorne effect. (142)

46. The notion that subjects are often motivated to "look good" when participating in an experiment is because they are
a. trying to be good subjects.
b. being employed as quasi-control subjects.
c. experiencing experimenter expectancy effects.
* d. experiencing evaluation apprehension. (142)

47. One strategy that researchers can use to identify potential demand characteristics in their experiment is to employ the use of
a. good subjects.
* b. quasi-control subjects.
c. expectancy control designs.
d. double-blind procedures. (143)

48. After conducting an experiment, Linda realizes that she smiled more often when interacting with the experimental group than when interacting with the control group. Which artifact may Linda have introduced into her experiment?
a. the double-blind procedure
b. the good subject effect
c. the Hawthorne effect
* d. the experimenter expectancy effect (143-144)

49. One way of minimizing the experimenter expectancy effect is to employ the use of _____ experimenters.
* a. blind
b. control
c. neutral
d. quasi (144)

50. A researcher used an expectancy control design in her study of the effectiveness of a new teaching method. Students were randomly assigned to conditions, and the following group means were found (where higher means reflect better performance):

	Teaching Method	
	New	Old
Expectation that method works	25	15
Expectation that method does not work	15	5

Assuming that all of the differences are statistically significant, what can the researcher conclude?
a. The old teaching method is more effective.
b. The condition with the expectation that the method does not work performed better.
* c. The new teaching method, combined with the expectation that the method works, performed the best.
d. Expectations did not have an impact on student performance. (143-144)

51. The expectancy control design is an example of which type of research design?
a. repeated measures design
b. quasi-experimental design
c. preexperimental design
* d. simple factorial design (143-145)

SHORT ESSAY QUESTIONS

1. How do randomized experiments differ from other types of experiments?

2. Describe three reasons why one should use random assignment of subjects to treatment conditions.

3. What is the difference between within-subjects and between-subjects research designs? Give an example of each type of design.

4. Why is counterbalancing important when using a within-subjects design with repeated treatments and measurements?

5. What are the four kinds of causation one may refer to when explaining events? Which type of causation do scientists refer to when they say that one thing "caused" something else?

6. Describe the three criteria for establishing efficient causation.

7. How do Mill's methods form the logical basis of all simple randomized designs?

8. What is a one-shot case study? Why is it inadequate for demonstrating efficient causation?

9. Describe six threats to internal validity identified by Campbell and his colleagues.

10. What is an artifact? Why are artifacts of concern to researchers?

11. What is the good subject effect? How can this effect be minimized in a study?

12. How can the experimenter expectancy effect bias the results of an experiment? How can this bias be minimized?

13. What is a quasi-control subject? How can this assist in discovering demand characteristics in an experiment?

CHAPTER 8: *NONRANDOMIZED RESEARCH AND CAUSAL REASONING*

CHAPTER OUTLINE

I. How is Causal Reasoning Attempted in the Absence of Randomization?
- A. Sometimes, for practical or ethical reasons, randomized experiments are not possible for teasing out causal relationships. These may resemble true experimental designs in that they have the equivalent of treatment conditions, outcome measures, and sampling units. However, they do not use randomization to allocate sampling units to treatment conditions.
- B. Several types of nonrandomized designs are frequently used for generalized causal inferences, including nonequivalent-groups designs, interrupted time-series designs, cross-lagged panel designs, single-case experimental designs, and cohort designs.
- C**. Prospective data** are data collected from an event or a point in time forward.
- D. **Retrospective data** are data collected backward in time.

II. How Is the Third-Variable Problem Relevant?
- A. The third-variable problem refers to nonrandomized studies' inability to conclusively rule out a rival hypothesis.

III. What Is Meant By Subclassification on Propensity Scores?
- A. **Nonequivalent-groups designs** are between-subjects designs in which the participants are assigned to experimental and control groups by means other than randomization and are tested before and after the experimental treatment.
- B. In some circumstances when researchers must work with intact groups, a "Type G Error" may exist between groups for which individual group members are not able to be randomly assigned to groups. That is, relevant extraneous factors may exist that are characteristic of one group but are uncharacteristic of the other group.
- C. One approach for trying to increase the likelihood that the two groups will be similar to one another is to **match** the groups based on the propensity score method.
- D. An innovative statistical way of improving this situation called **subclassification on propensity scores** can be used when sample sizes are large enough and there are relevant subgroups that are also well stocked with sampling units. This procedure reduces all of the variables on which the "treated" and "untreated" sampling units differ to a single composite variable, called a **propensity score**.

IV. What Are Time-Series Designs and "Found Experiments"?
- A. The defining characteristic of **time-series designs** is the study of variation across some dimension over time.
  1. The data structure for these designs is called a time-series because there is a single data point for each point in time.
  2. It is frequently called an "interrupted time-series" when there is a clear dividing line at the beginning of the intervention.

B. The sociologist David P. Phillips has reported a number of applications of this approach, calling them "found experiments" because they are essentially quasi-experiments in naturally occurring situations.

V. What Within-Subjects Designs Are Used in Single-Case Experiments?

A. A very popular class of within-subjects designs is called **single-case experimental research.** These designs are also often referred to as **small-*N* experimental research** and ***N*-of-1 experimental research**.

B. Generally speaking, all single-case experiments have the following characteristics:

1. Only one sampling unit is typically studied or only a few units are studied.
2. Repeated measures are taken of the unit (a within-subjects design).
3. Random assignment is rarely used.

C. Single-case experimental designs are particularly popular in educational, clinical, and counseling settings for evaluating the effects of operant conditioning interventions.

1. Such designs use a **behavioral baseline** to serve as a kind of "pretest" against which the pattern of behavior exhibited after the treatment has been administered can be compared.
2. In other words, the unit serves as its own control group in a simple within-subjects design.

D. The basic model is called an **A-B-A design**, which evolved out of an even simpler prototype, the **A-B design** (which is the simplest of all single-case designs).

E. A number of other single-case designs are used in clinical intervention assessment.

1. The **A-B-BC-B design** where B and C are two different therapeutic interventions.
2. The **A-B-A-B design** provides two occasions for demonstrating the positive effects of the intervention.

VI. How Are Correlations Interpreted in Cross-Lagged Panel Designs?

A. A **cross-lagged panel design** is called cross-lagged because some data points are treated as temporarily "lagged" (or delayed) values of the outcome variable. It is called a panel design because a panel study is another name for a **longitudinal study**.

B. The basic premise of this design is that if the **test-retest correlations** are about equally reliable, and the **synchronous correlations** are about equal in magnitude, then the **cross-lagged correlations** between the two sets of data points should suggest the more likely causal direction, or which variable shows the preponderance of causal influence.

VII. What Is the Difference Between Longitudinal and Cross-Sectional Research?

A. Sometimes researchers are interested in studying the life course of some variable of interest.

B. **Cross-sectional designs** are designs that take a slice of time and compare subjects on one or more variables simultaneously. A problem with these designs is that there is a possible confounding of **cohort** and maturation.

C. Researchers who like to use longitudinal designs also attempt, whenever possible, to examine several cohorts cross-sectionally and longitudinally. This allows them to learn about cohort changes as well as age group changes as a function of period.

LECTURE IDEAS AND ACTIVITIES

1. One way to introduce the different quasi-experimental designs is to have the class discuss how one can investigate research questions which, for either practical or ethical reasons, cannot be investigated using true experimental designs. For example, ask students whether it is appropriate to use a true experimental design to investigate the impact that parenting practices may have on the development of a child's social skills. If students conclude that such a study would be unethical, ask them whether such research questions should be abandoned. This can start a discussion of how quasi-experimental designs can be used to answer such research questions.

2. A controversial issue in society is whether exposure to second-hand smoke increases one's risk of getting lung cancer. Have students develop possible quasi-experimental designs that can be used to evaluate the veracity of this claim. For example, an investigator might use an interrupted time-series to evaluate the incidence of lung cancer among nonsmokers before and after the implementation of laws banning smoking in public places. Discuss how one can use Mill's methods for evaluating causality based on the circumstantial evidence surrounding this issue.

3. James Carr and John Austin (1999) introduce single-subject research designs in their class by having their students conduct an *N-of-1* study where they are the subjects. In this study, the students are told that they will be investigating the effects of exercise on one's heart rate using an A-B-A design. During the baseline phase, students are instructed to record their pulse for a one-minute period. After recording their pulse rate, the students again measure and record their pulse rate for four more one-minute periods. Upon completion of the baseline period, students are instructed that they will be exercising for 20 seconds (Carr and Austin have their students do jumping jacks), after which they will sit and measure their pulse rate for a one-minute period. This exercise-recording process lasts for five treatment periods. Once the exercise phase of the study has ended, the students once again measure and record the pulse rate during five one-minute periods. Upon completion of the study, Carr and Austin have their students create a line graph of the 15 data points so that their students may learn how to analyze the results from single-subject research designs in order to make causal inferences.

Carr, J. E., & Austin, J. (1999). A classroom demonstration of single-subject research designs. In M. E. Ware and C. L. Brewer (Eds.), *Handbook for teaching statistics and research methods (2nd ed.)* (pp. 215–218). Mahwah, NJ: Lawrence Erlbaum.

4. Consistent with Rosnow and Rosenthal's discussion of the importance of methodological pluralism, Kenneth Kerber (1996) has his students study the same phenomenon using both a correlational and experimental study to show how they are complementary approaches. Kerber describes specifically the correlational and experimental studies his students conducted to explore the impact of costs and rewards on helping, but Kerber's approach is applicable to investigation of almost any psychological phenomenon of interest.

Kerber, K. W. (1996). Rewards, costs, and helping: A demonstration of the complementary nature of experimental and correlational research. In M. E. Ware and D. E. Johnson, (Eds.), *Handbook of demonstrations and activities in the teaching of psychology* (Vol. 1, pp. 148–150). Mahwah, NJ: Lawrence Erlbaum.

MULTIPLE-CHOICE QUESTIONS

1. An educator wants to evaluate the efficacy of a new teaching technique. He uses his class as the experimental treatment group and a colleague's class as the control group. Which type of experimental design is the educator most likely employing?
   a. correlational research design
   * b. nonequivalent-groups design
   c. cross-lagged panel design
   d. interrupted time-series design (152)

2. In a study of the death rates of smokers and nonsmokers in three countries, it is found that the nonsmokers have higher death rates than the smokers. What technique could be used to address difficulties with the sampling units that produced this unusual result?
   a. There is no technique available to address it
   * b. subclassification on propensity scores
   c. time series design
   d. lagged correlations (153)

3. In situations where one must work with intact groups but where there are also ethical implications concerning depriving both groups of the benefits of the experimental treatment, one uses a randomized design with a _____ control group.
   a. zero
   b. preexperimental
   * c. wait-list
   d. placebo (153)

4. Which of the following is NOT a design that involves comparisons of outcome measures obtained at different time intervals before and after the introduction of the treatment?
   * a. nonequivalent-groups design
   b. single-case experimental designs
   c. cross-lagged panel designs
   d. time-series designs (152)

5. To determine whether a new anti-drug campaign has had an impact on students' usage of drugs, an educator compares students' drug usage during the five weeks before and the five weeks after the anti-drug campaign. Which quasi-experimental design characterizes the educator's investigation?
   a. single-case experimental design
   b. cross-lagged panel design
   * c. time-series design
   d. nonequivalent-groups design (154)

6. Fred systematically observes how changes in lighting affect his rat's lever-pressing behavior. Which research design best characterizes Fred's investigation?
   a. cross-lagged panel design
   * b. single-case experimental design
   c. nonequivalent groups design
   d. correlational design (155)

7. Which of the following is NOT a characteristic of single-case experimental designs?
   a. the use of a within-subjects design
   b. studying one or only a few sampling units
   c. the manipulation of the independent variable
   * d. the random assignment of sampling units to treatment conditions (155)

8. The observations of an individual's behavior before the experimental treatment in a single-case experimental design is referred to as a(n)
   a. pretest.
   * b. behavioral baseline.
   c. experimental treatment.
   d. cohort. (156)

9. To evaluate the efficacy of a new intervention procedure, a clinician observes a patient's behavior before and after the intervention. The clinician then repeats the procedure, again observing the patient's behavior before and after the introduction of the treatment. Which single-case experimental design best describes the clinician's investigation?
   * a. A-B-A-B design
   b. A-B-BC-B design
   c. A-B-C-A design
   d. A-B-AB-A design (157)

10. A professor is interested in whether one's confidence influences one's self-esteem or whether it is one's self-esteem that influences one's self-confidence. He measures his students' levels of self-esteem and confidence and learns that they are positively correlated. One year later the professor reassesses the students' self-esteem and confidence levels. He finds that there is a stronger correlation between one's confidence at time one and one's self-esteem one year later than the correlation between one's self-esteem at time one and one's confidence level one year later. Which type of quasi-experimental design is the professor employing?
    a. a single-case experimental design
    b. a time-series design
  * c. a cross-lagged panel design
    d. nonequivalent groups design (158)

11. Which type of study examines changes in a person or group of people over an extended period of time?
    a. cohort
    b. preexperimental
  * c. longitudinal
    d. historical control (158)

12. Another name for a longitudinal study is a(n)
  * a. panel study.
    b. nonequivalent groups design.
    c. interrupted time-series.
    d. preexperimental design. (158)

13. In a cross-lagged panel design, which correlation indicates the reliability of the same variables over time?
  * a. test-retest
    b. synchronous
    c. cross-lagged
    d. causality (158)

14. In a cross-lagged panel design, which correlation can be used to compare the reliability of the association between two variables measured at the same time?
    a. causal correlation
    b. cross-lagged correlation
  * c. synchronous correlation
    d. test-retest correlation (158-159)

15. In a cross-lagged panel design, which correlation is used to infer a causal relationship?
    a. synchronous correlation
    b. causal correlation
    c. test-retest correlation
  * d. cross-lagged correlation (158-159)

16. Which design examines several age groups during one slice of a time period?
    a. longitudinal
    b. cross-lagged

* c. cross-sectional
d. cohort (160)

17. June wants to know if one's interest in sports declines as one gets older. She compares several age groups with respect to how much they like to watch sports on television. Which design best describes June's study?
* a. cross-sectional
b. nonequivalent-groups
c. cohort
d. longitudinal (160)

18. A group of individuals who are born and grow up in the same period and thus have generally similar life experiences are referred to by researchers as a
a. family.
b. collective.
* c. cohort.
d. nonequivalent group. (160)

19. Ralph wants to examine group-related changes as well as age-related changes as a function of time. Which quasi-experimental design would be most appropriate for him to use?
a. nonequivalent-groups design
b. time-series design
* c. longitudinal design using cohorts
d. single-case experimental design (160)

SHORT ESSAY QUESTIONS

1. Why would one sometimes use a nonrandomized design?

2. What is a nonequivalent-groups design? Describe one way that an investigator can try to minimize the threat to internal validity that is inherent in this type of design.

3. What is the difference between prospective and retrospective data?

4. What is the third-variable problem? How is this relevant to nonrandomized studies?

5. What is subclassification on propensity scores and how can it be used to analyze data to better address nonequivalent-groups designs?

6. What is a time-series design? Describe a research question that this design can be used to address.

7. What are three characteristics of single-case experimental designs?

8. Differentiate between an A-B-A design, an A-B-BC-B design, and an A-B-A-B design.

9. What is a cross-lagged panel design? How can this design be used to make inferences of causation?

10. What is a problem with using cross-sectional designs to investigate the life-course of some variable of interest? How can longitudinal designs using cohorts minimize this problem?

CHAPTER 9: *SURVEY RESEARCH AND SUBJECT RECRUITMEN*T

CHAPTER OUTLINE

I. What Are Opportunity and Probability Samples?
- A. When experimenters in behavioral science are interested in learning about human nature in general, they typically use **opportunity samples**, samples made up of the first units available.
- B. Most survey researchers are interested in generalizing their results to a specified larger pool (or **population**) of individuals.
    1. Using an opportunity sample can produce spurious results and misleading conclusions.
    2. Survey researchers recruit participants based on sampling lists (or frames) that identify the relevant units or subgroups in the population.
- C. Survey research focuses on a segment (or **sample**) that is believed to be typical of the population rather than questioning every member of the population (which is usually impossible).
- D. All survey studies use **sampling plans** in which some method of **probability sampling** determines the **random selection** of subjects.
- E. Sampling plans enable the researcher to assume reasonably – but with no guarantee of being correct – that the sample is **representative** of its population.

II. What Is Meant By Bias and Instability in Survey Research?
- A. There are two important statistical requirements of a probability sampling plan.
    1. The sample values must be **unbiased** in that the values of the sample must, on the average, coincide with the "true values" of the population. The distance between the true population value and the midpoint of the sampling units indicates the amount of **bias** (i.e., systematic error).
    2. There must be **stability** in the samples, meaning that there is not much variability (or spread) in the sample values.
- B. Generally speaking, the more homogeneous or alike the members of the population are, the fewer members are needed to be sampled. The more heterogeneous (dissimilar) the members are, the more members are needed in order to ensure that the full range of dissimilarity is represented in the sample.

III. Why Do We Not Know "For Sure" the Bias in Sampling?
- A. One way to know for sure if the sample is unbiased, is to examine every single member of the population and the sample *at the same time* the sampling is done.
- B. Election forecasting is sometimes thought to allow us to detect bias in a sample because we can compare the predicted voting results with the actual votes; however, we are comparing data obtained at one point in time with the results at another point in time.

IV. How Is Simple Random Sampling Done?
- A. The basic prototype of probability sampling is called **simple random sampling**.

1. This sample plan is considered to be "simple" in that the sample is selected from an undivided population.
2. This plan is considered to be "random" in that each sampling unit has an equal chance of being selected and that the selection of any one unit has no influence on the selection of other units in the population.

B. A further requirement for simple random sampling is that one must have knowledge of the existence of all units in the population.

C. There are a variety of options for randomly selecting units including **random digit dialing** when conducting telephone interviews and the use of a table of random digits.

D. **Sampling with replacement** means that the selected names are placed in the selection pool again and may be reselected on subsequent draws.

E. In **sampling without replacement**, a previously selected name cannot be reselected and must be disregarded on any later draw.

F. Either option is technically acceptable, but survey researchers generally prefer sampling without replacement as they do not want to use the same sampling units twice or more.

V. What Are Stratified Random Sampling and Area Probability Sampling?

A. Simple random sampling is frequently used when the population is known to be homogeneous or its precise composition is unknown.

B. When something is known about the exact composition of the population, a more efficient sampling plan involves sampling from the different substrates of the population (called **strata** or **clusters**).

C. In **stratified random sampling**, a separate sample is randomly selected from each homogenous stratum (or "layer") of the population. The stratum means are then statistically weighted to form a combined estimate for the entire population.

D. **Area probability sampling** involves dividing the population into subpopulations based on meaningful geographic areas rather than some characteristic of the population.

E. The assumption is that, within each of the areas, the sampling units will have the same probability of being chosen.

F. Although complicated, the method is cost-effective because the research design can be used repeatedly with only minor modifications.

G. The key requirements for this plan are that:

1. All areas will have some chance of selection.
2. Units within the areas are chosen impartially.

VI. What Did the *Literary Digest* Case Teach Pollsters?

A. The flawed presidential prediction made by the magazine, the *Literary Digest*, demonstrated that large numbers do not, in and of themselves, increase the representativeness of the population. Gallup's predictions, using **quota sampling**, an early precursor of current methods, correctly forecasted the presidential results.

B. Attention must be given to the factor of voter turnout when making election predictions.

C. Political polling should be done as close to Election Day as possible.

VII. What Are Point Estimates and Interval Estimates?

A. **Point estimates** are designed to indicate some particular characteristic of the population.

B. **Interval estimates** indicate how much the point estimates are likely to be in error.

C. The **confidence interval** indicates the probability that the estimated population value is correct within a plus-or-minus specified interval.

D. To obtain an approximate 95% confidence interval around the proportion (symbolized as *prop*):

$$\text{Lower Limit} = prop - 2\sqrt{\frac{prop(1-prop)}{n}}$$

and

$$\text{Upper Limit} = prop + 2\sqrt{\frac{prop(1-prop)}{n}}$$

VIII. What Are the Benefits of Stratification?

A. Although both simple and stratified random sampling are considered to be **unbiased sampling plans**, there are advantages to using stratified random sampling.

B. Stratified random sampling reduces the **error of the estimate**.

IX. How Is Nonresponse Bias Handled in Survey Research?

A. One difficulty in obtaining random samples is the refusal of some individuals to participate in survey research.

B. **Nonresponse bias** is the error due to nonparticipation or nonresponse.

1. This bias can result in a smaller **effective sample size** than the researcher planned on for statistical reasons.
2. This bias can also seriously jeopardize the accuracy of estimates of population values.

C. Having information on all members of a population can be used to compare respondents to nonrespondents.

D. Increasing response rate may reduce nonresponse bias.

X. What Are the Typical Characteristics of Volunteer Subjects?

A. There are several reasons why most experimenters do not concern themselves with the particulars of a probability sampling plan when recruiting participants for experimental and quasi-experimental studies.

1. Often it is impossible to work within the confines of such a plan.
2. Even when random subject selection is possible, many experimenters believe that "people are people" in terms of the psychological factors or mechanisms they are studying.

B. However, it has been found that the use of volunteers may lead to biased conclusions, even when the volunteers were randomly assigned to the experimental and control conditions.

C. To explain **volunteer bias** (and to explain how it is controlled), one must know the characteristics of typical volunteers and nonvolunteers.

1. The text lists nine characteristics of those more likely to volunteer to participate in a research study.

2. The degree of bias these volunteer characteristics may introduce depends, to a large degree, on the context or nature of the research study.

XI. How Is Volunteer Bias in Opportunity Samples Managed?

A. To address volunteer bias, researchers can attempt to stimulate participation by typical nonvolunteers. Some techniques include:
   1. Explain why the research is being done and why subjects will find the research worthwhile.
   2. Avoid evaluation apprehension.
   3. Emphasize importance of the research.
   4. Offer incentives.
   5. Avoid stressful procedures.
   6. Treat recruits like a "granting agency" (they grant your study their time and effort).

B. Whatever your research project, whether it involves a survey, a randomized experiment or another strategy of collecting data directly from people, the final step before implementing the study is to pilot-test the materials.

LECTURE IDEAS AND ACTIVITIES

1. To emphasize human limitations in generating random as opposed to aimless sequences in numbers, use the discussion from Box 9.2 in the text as a class exercise. Have each student write down "at random" ten single-digit numbers ranging from 0 to 9. Collect the numbers from all of the students, tabulate and graph on the board the 0s, 1s, 2s, etc. What will become obvious is that each of the digits will not occur approximately 10% of the time (if they were truly randomly selected), but rather some digits will be overrepresented in this "random selection" of digits. Discuss these results in the context of the difference between randomness and aimlessness.

2. Louis Snellgrove (1981) suggests an exercise to demonstrate the basic concepts of sampling. Snellgrove places 500 BBs into an adequate size container. These BBs are painted red, blue, green, yellow, or natural such that there are 100 BBs of each color. (NOTE: beads, buttons, or other small objects can be used in place of BBs.) The container of BBs is then shaken vigorously to "mix up" the BBs. The container of BBs represents the "population" and students are asked to draw BBs from the container in order to form a sample from which they can estimate the color composition of the BBs in the container. You can have students randomly select different sample sizes of BBs in order to demonstrate that the larger the sample, the more it becomes representative of the population. Based on this demonstration, you may want to begin a discussion of how national surveys can use a random sample of 1500 people to accurately estimate the responses of 200 million people.

Snellgrove, L. (1981). Sampling and probability. In L. T. Benjamin, Jr. and K. D. Lowman (Eds.), *Activities handbook for the teaching of psychology* (Vol. 1, pp. 12–13). Washington, DC: American Psychological Association.

3. Randolph Smith employs a "tasty" way of introducing his students to the concepts of sampling. Smith distributes fun packs of plain M&M candy pieces. The students are to examine their "data" (i.e., candy pieces), making a simple frequency distribution of the six M&M colors contained in their "sample" (i.e., candy package). Students are then asked to develop a hypothesis based on their sample about the distribution of the different colors of plain M&M candies in the population (i.e., make point estimates). Students are then asked to combine their frequency distribution with another student and generate a joint hypothesis. Finally, Smith has the entire class pool their results to generate an overall hypothesis of the distribution of colors in the population. Through this process of combining their data with others, students can see the impact that sample size has on the variability of their point estimates. While the official point estimates for all M&M products are no longer available at the official M&M website (http://www.m-ms.com), in the past the point estimates for the six colors were as follows: 30% brown; 20% yellow; 20% red; 10% green; 10% blue; and 10% orange.

Smith, R. A. (1999). A tasty sampler®: Teaching about sampling using M&M's. In L. T. Benjamin, B. F. Nodine, R. M. Ernst and C. B. Broeker (Eds.), A*ctivities handbook for the teaching of psychology* (Vol. 4, pp. 66–67). Washington, DC: American Psychological Association.

MULTIPLE-CHOICE QUESTIONS

1. Samples that consist of individuals who are easily available to the researcher are referred to as
   a. random samples.
   * b. opportunity samples.
   c. volunteer samples.
   d. biased samples. (164)

2. To describe the characteristics of a population, survey researchers rely on responses from individuals from the population who
   a. have volunteered to give the information.
   b. have been randomly assigned to the population.
   * c. have been randomly selected from the population.
   d. are knowledgeable about the population. (168)

3. Survey researchers collect information from a _____ to infer "how things are" in a _____.
   a. population; sample
   * b. sample; population
   c. cohort; sample
   d. sample; cohort (164)

4. When conducting a survey, one often uses a _____ to learn about the _____ of interest.
   a. population; sample
   b. control group; population

c. subset; sample
* d. sample; population (164)

5. A major concern survey researchers have about samples is whether or not the samples are _____ the population.
a. familiar with
b. knowledgeable about
c. identical to
* d. representative of (165)

6. To increase confidence that a sample is representative of the population, survey researchers utilize _____ in order to select the sample.
a. randomized assignment plans
* b. probability sampling plans
c. stratified random assignment plans
d. probability assignment plans (165)

7. Another name for systematic error is
a. stability.
b. strata.
* c. bias.
d. point estimate. (165-166)

8. Values from a sample that, on average, coincide with the "true" values of the population are referred to as _____ values.
a. precise
b. accurate
c. stable
* d. unbiased (166)

9. The true population mean is 4. Sample A has the following values: 3, 4, 4, 5. Sample B has the following values: 0, 4, 4, 8. Compared to Sample B, Sample A is more
a. unbiased.
b. biased.
* c. stable.
d. unstable. (166)

10. In order to estimate the number of hours a day the typical student studies at his university, Paul interviews 25 randomly selected pre-med students. Paul has most likely violated which requirement for probability sampling plans?
a. stability in the samples
* b. unbiased sample values
c. inadequate sample size
d. the use of random selection (166)

11. Interested in the number of hours a day the typical student studies at his university, Paul decides to form several samples of students and interview each sample at a different time during the semester, including during the final exam period. Which requirement for probability sampling plans should Paul be especially concerned that he may be violating?
    a. the use of opportunistic samples
    b. unbiased sample values
  * c. stability in sample estimates
    d. inadequate sample sizes (166)

12. Which probability sampling plan is used when the population is undivided?
  * a. simple random sampling
    b. stratified random sampling
    c. area probability sampling
    d. cluster random sampling (168)

13. In order to prevent sampling units from being measured more than one time, survey researchers often use
    a. unbiased sampling.
  * b. sampling without replacement.
    c. sampling with replacement.
    d. stratified sampling with replacement. (170)

14. A probability sampling plan in which each unit continues to have the same exact probability of being sampled every time a unit is chosen is an example of
    a. unbiased sampling.
    b. sampling without replacement.
  * c. sampling with replacement.
    d. stratified sampling without replacement. (170)

15. Subpopulations that are created by dividing a population into smaller groups are often referred to as
    a. samples.
  * b. strata.
    c. quasi-populations.
    d. simple populations. (170-171)

16. To survey the student body, Bill separates students according to their year in school and then randomly selects students to interview from each group. Which sampling plan is Bill utilizing?
    * a. stratified random sampling
    b. area probability sampling
    c. cohort probability sampling
    d. simple random sampling (171)

17. To assess living conditions on campus, Kim groups students according to the dormitories where they live and then randomly selects students from each dormitory to interview. Which of the following best describes Kim's sampling plan?
    a. simple random sampling
    b. stratified random sampling
    c. cohort probability sampling
    * d. area probability sampling (171)

18. After surveying a randomly selected sample of students, Bill concludes that students at his school study an average of two hours a day. Bill's conclusion is an example of a(n)
    a. biased estimate.
    * b. point estimate.
    c. interval estimate.
    d. sample estimate. (173)

19. Bill concludes that he is 95% confident in his conclusion that the typical university student studies between 2 and 4 hours a day. This is an example of a(n)
    a. point estimate.
    b. sample estimate.
    c. inaccurate estimate.
    * d. interval estimate. (173)

20. Which of the following refers to how good one's estimate is based on an unbiased sampling plan?
    a. bias estimate
    b. point estimate
    * c. error of estimate
    d. stability estimate (174-175)

21. Jerry is concerned that the individuals who did not return his survey might somehow be different from those who did return the survey. Jerry is concerned about the possibility of a(n)
    a. sample stability problem.
    * b. nonresponse bias.
    c. volunteer bias.
    d. inadequate sample. (176)

22. Anne is worried because the only students who are willing to participate in her study are those who consistently make the Dean's List. Anne is most likely concerned about the
    a. nonresponse bias.
    b. stability of her sample.
    c. error of estimate in her sample.
* d. volunteer bias. (179)

SHORT ESSAY QUESTIONS

1. What is the difference between random selection and random assignment?

2. Describe the two requirements for a probability sampling plan.

3. What is the difference between sampling with replacement and sampling without replacement?

4. A student employs a sampling plan that involves randomly selecting a name from a list and then selecting every third person thereafter. Why can this plan not be considered a simple random sampling plan?

5. What is stratified random sampling? What is area probability sampling? What advantages do these sampling plans have over simple random sampling?

6. Describe one methodological lesson that George Gallup learned about conducting survey research.

7. What is the difference between point estimates and interval estimates?

8. What is meant by the nonresponse bias? How might this bias be minimized?

9. What is meant by the volunteer subject problem? Why is it of concern to researchers?

10. Describe three ways that volunteers for research have been found to differ from nonvolunteers.

11. How might one minimize the volunteer subject problem when recruiting research participants?

12. Why should pilot-testing be utilized when determining the procedures to be used in the selection of research participants?

CHAPTER 10: *SUMMARIZING THE DATA*

CHAPTER OUTLINE

I. How Is Visual Integrity Ensured When Results Are Graphed?
- A. Properly done visual displays are informative and easy to understand.
- B. Edward R. Tufte (1983) suggests three criteria for visual integrity:
  1. Clarity, representing data closely integrated with their numerical meaning.
  2. Precision, exact and not exaggerated representation.
  3. Efficiency, using compact space and avoiding distractions.
- C. Cognitive psychologist Stephen M. Kosslyn (1994) suggests data presentations keep in mind three interrelated principles regarding how the mind processes information:
  1. The mind is not a camera.
  2. The mind judges a book by its cover.
  3. The spirit is willing, but the mind is weak.

II. How Are Frequencies Displayed in Tables, Bar Graphs, and Line Graphs?
- A. **Frequency distributions** are useful for emphasizing the overall pattern of the data.
- B. **Bar graphs** are especially useful for representing categories of responses and frequencies (or proportions) within those categories.
- C. **Line graphs** are an efficient way of graphing changes in the frequency (or proportion) of scores over time.

III. How Do Stem-and-Leaf Charts Work?
- A. A **stem-and-leaf chart** is a hybrid between a table and a graph because it presents the original numbers and simultaneously gives an economic summary view of them.
- B. The beauty of the stem-and-leaf chart is that it allows one to see the data as a whole and to notice:
  1. How symmetrical the data set is.
  2. How spread out the scores are.
  3. Whether there are small and large concentrations of scores.
  4. Whether there are any gaps in the scores.
- C. **Back-to-back stem-and-leaf charts** are useful for comparing data from two different sets of scores.

IV. How Are Percentiles Used to Summarize Part of a Batch?
- A. There is often a practical value in knowing the point in the distribution below and above which a certain percentage of sampling units fall. This point is called the **percentile**.
- B. The 50^{th} percentile is also called the **median** (symbolized as *Mdn*).
  1. The median is one measure of **central tendency**. It is the midmost score in the distribution of scores.
  2. To locate the median, the value $.50(N + 1)$ (where N is the total number of scores in the ordered set) indicates the score in the ordered set that represents the median.

3. Similarly, the 75th percentile is .75(N + 1), the 25th percentile is .25(N + 1), etc.
4. The distance between the 25th and the 75th percentiles is called the **interquartile range**.

V. How Is an Exploratory Data Analysis Done?

A. The stem-and-leaf can be used not only to do hypothesis testing (known as confirmatory data analysis), but to do exploratory data analysis as well.

B. The minimum and maximum values, the 25th and 75th percentiles, the interquartile range, and the median all help to characterize the nature of a batch of data.

VI. How Does Asymmetry Affect Measures of Central Tendency

A. The **mode**, another measure of central tendency, is the score, or category of scores, that occurs with the greatest frequency. A data set with two modes is **bimodal**.

B. A third measure of central tendency is the **arithmetic mean**, or **mean**, which is symbolized as M (although another symbol sometimes used is $\overline{X}$). The formula for the mean is

$$M = \frac{\sum X}{N}$$

C. Reporting more than one measure of central tendency provides a clearer picture of the distribution of the data set.

1. The median, mean, and mode are equal in a **symmetrical distribution**.
2. In a **positively skewed distribution**, the mean of the distribution is much larger than the median.
3. In a **negatively skewed distribution**, the mean of the distribution is smaller than the median.
4. The mean is the measure of central tendency most affected by outliers.

D. Scores that lie far outside the normal range are called **outliers**.

E. When a distribution is strongly asymmetrical because of outliers, researchers frequently prefer a **trimmed mean** to an ordinary mean.

F. Until you know whether the outliers are errors (scoring or recording mistakes), dropping them could be discarding important information that warrants further examination. There are also statistical procedures for "reeling in" outliers and making them part of the group.

VII. How Do I Measure How "Spread Out" a Set of Scores Is?

A. Besides knowing the central tendency (or "typical value") of a set of scores, researchers also usually want to know how "spread out" the scores are.

B. There are different measures of spread, dispersion, or variability such as the interquartile range, which indicates the variability characteristic of the middle 50% of the scores.

C. The ordinary **range** (or **crude range**) is the difference between the highest and lowest scores.

1. How the range is interpreted depends on the purpose or objective of the study and the nature of the instruments used.
2. The **extended range** (sometimes called the **corrected range**) adds a half unit at the top of the distribution and a half unit at the bottom of the distribution when

calculating the range in order to account for the lack of precision in the measurement instrument.

D. The **variance** (or **mean square**) is the mean of the squared deviations of the scores from their mean.

1. The symbol used to denote the variance of a population is σ^2 ("sigma squared").
2. The formula used to calculate the population variance is:

$$\sigma^2 = \frac{\sum (X - M)^2}{N}$$

E. The **standard deviation**, the most widely used and reported measure of spread around the average, is simply the square root of the population variance.

1. The standard deviation of a population is symbolized as σ.
2. The formula for the standard deviation is:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum (X - M)^2}{N}}$$

3. Another name for the standard deviation is the **root mean square**.

VIII. What Are Descriptive and Inferential Measures?

A. **Descriptive measures** are used when one is describing the complete population of scores or events. Greek letters are used to symbolize these measures.

B. **Inferential measures** are used when one wants to generalize from a sample of known scores or events to a population of unknown scores or events, which may be finite or infinite. Inferential measures are symbolized with roman letters.

C. Except for the denominator and the symbol, the descriptive and inferential formulas for computing variances (σ^2 and S^2) are identical. Similarly, the same is true for the standard deviation (σ and S).

1. In the descriptive formulas for variances and standard deviations, the numerator is divided by N.
2. In the inferential formulas for variances and standard deviations, the numerator is divided by $N - 1$. The resulting statistic for sample variance (i.e., S^2) is referred to as the **unbiased estimator of the population value of σ^2**.

IX. How Do I Estimate a Confidence Interval (CI) Around a Population Mean?

A. The most commonly reported confidence interval is the 95% CI.

1. Three quantities are required to compute a 95% CI around an obtained estimate of a population mean: N, S, and $t_{.05}$.
2. To obtain a 95% confidence interval around the estimated population mean:

$$\text{Lower Limit} = M - \frac{(t_{.05})(S)}{\sqrt{N}}$$

and

$$\text{Upper Limit} = M + \frac{(t_{.05})(S)}{\sqrt{N}}$$

B. Confidence intervals tell us how accurately we have estimated a quantity.

C. The 2010 American Psychological Association (APA) Publication Manual recommends that it is best to use a single confidence level specified on an a priori basis throughout a manuscript.

X. What Is Distinctive About the Normal Distribution?

A. The distribution of scores collected by means of a representative sampling procedure will often form a curve that has a distinct bell-like shape. This curve is called a **normal distribution**.

B. The normal distribution is useful for providing a mathematical description for populations because it can be completely described from just the mean and the standard deviation.

C. The normal distribution enables one to translate raw scores into standard deviation units, allowing one to compare scores from different data sets.

D. Statistics derived from the normal distribution are important in the testing of the hypothesis.

XI. Why Are *z Scores* Called Standard Scores, and How Are They Used?

A. A normal curve with a mean set equal to 0 and a standard deviation set equal to 1 is called a **standard normal curve**.

B. Any individual raw score can be statistically translated (referred to as **transformation**) into a **standard score** corresponding to a location on the abscissa of a standard normal curve.

1. A standard score (or ***z score***) expresses, in standard deviation units, the raw score's distance from the mean of the normative group.
2. The formula for a standard score is:

$$z \text{ score} = \frac{X - M}{\sigma}$$

C. By converting an individual raw score into a *z* score, one can determine the proportion of scores that includes and is higher or lower than that score in the normal distribution using the Table of *z* Values and Their Associated One-Tailed *p* Values (Standardized Normal Deviates) (Table B.1 in Appendix B).

LECTURE IDEAS AND ACTIVITIES

1. One major challenge you may face with this chapter and subsequent chapters is "statisticophobia" (Dillon, 1996) or fear of statistics. To help students overcome this fear, you may want to reiterate the point made in the text that statistics are tools that researchers use to evaluate and make sense of research data. The goal for students is to feel comfortable enough with statistics so that they are not only able to analyze and report their own research results, but so that they are also able to critically evaluate the statistical claims made by other researchers.

Frances Conners, Steven McCown, and Beverly Roskos-Ewoldsen (1998) argue that instructors face four unique challenges when teaching a statistics course. These challenges include motivating students, minimizing math anxiety among students, handling performance extremes common among students in statistics classes, and making learning last. Conners et al. discuss several strategies for meeting these challenges. For example, students may be motivated if the instructor involves active as well as mastery learning into the course. Math anxiety might

be minimized by providing tutoring resources and relieving exam pressures through unlimited time and allowing repeat examinations. Conners et al. suggest that performance extremes could be handled through the use of peer tutoring in the classroom as well as an increased reliance on concrete presentations of the lecture material. They also recommend taking a proactive approach to addressing the needs of both strong and weak students. Finally, to help make learning last, Conners et al. suggest that instructors should focus on helping students achieve initial understanding of the material. In addition, they suggested that instructors should focus on providing students with memory cues to aid in recall of the material as well as help students understand common misperceptions in statistics.

Conners, F. A., McCown, S. M., & Roskos-Ewoldsen, B. (1998). Unique challenges in teaching undergraduate statistics. *Teaching of Psychology, 25,* 40–42.
Dillon, K. M. (1996). Statisticophobia. In M. E. Ware and D. E. Johnson (Eds.), *Handbook of demonstrations and activities in the teaching of psychology: Volume I* (p. 65). Mahwah, NJ: Lawrence Erlbaum.

2. A problem that you may face is that students do not immediately see how statistics can be used to answer "real life" questions. Rather than using fictional data sets, it might be useful to develop example data sets using "real world" data in order to illustrate the statistical concepts discussed in this and subsequent chapters. For example, Mark Shatz (1988) used a strike by Greyhound bus drivers as an opportunity to introduce the use of descriptive statistics. Students were provided fictional salary data designed to represent the circumstances involved in the Greyhound labor dispute. Students were asked to compute and interpret the descriptive statistics for this data set. Shatz reported that this relevant, albeit fictional, data set helped students not only to calculate descriptive statistical measures, but also to understand how descriptive statistics can be used to answer research questions.

Betsy Levonian Morgan (2001) has her students create "real world" or "genuine" data sets by extracting information from obituaries (e.g., age at death, deceased number of children, gender, etc.). Because reading obituaries may be upsetting to some students, you might also want to use a strategy employed by Steven Stern (1999) for developing "real world" data sets (see also Hettich, 1988; Thompson, 1999). During the first week of class, Stern has his students survey ten males and ten females, inquiring about the number of pairs of shoes they own. Students then use this information to learn how to visually and quantitatively summarize their data. For example, students can create back-to-back stem-and-leaf charts comparing the number of pairs of shoes owned by men and women in the sample. Because the potential is high for having outliers in this data set, the importance of evaluating the symmetry of the distribution can be discussed as well as potential methods for handling outliers (e.g., using the trimmed mean rather than the arithmetic mean). This data set can also be used when discussing hypothesis testing such as *t*-tests (see Chapter 13).

Hettich, P. (1988). The student as data generator. In M. E. Ware and C. L Brewer (Eds.), *Handbook for teaching statistics and research methods* (pp. 60–61). Hillsdale, NJ: Lawrence Erlbaum.
Morgan, B. L. (2001). Statistically lively uses for obituaries. *Teaching of Psychology, 28,* 56–58.

Shatz, M. A. (1988). The Greyhound strike: Using a labor dispute to teach descriptive statistics. In M. E. Ware and D. E. Johnson (Eds.), *Handbook of demonstrations and activities in the teaching of psychology: Volume I* (pp. 73–74). Mahwah, NJ: Lawrence Erlbaum.

Stern, S. E. (1999). The effect of gender on the number of shoes owned: Gathering data for statistical and methodological demonstrations. In L. T. Benjamin, B. F. Nodine, R. M. Ernst and C. B. Broeker (Eds.), *Activities handbook for the teaching of psychology* (Vol. 4, pp. 74–76). Washington, DC: American Psychological Association.

Thompson, W. B. (1999). Making data analysis realistic: Incorporating research into statistics courses. In M. E. Ware and C. L. Brewer (Eds.), *Handbook for teaching statistics and research methods (2nd ed.)* (pp. 3–6). Mahwah, NJ: Lawrence Erlbaum.

3. To help students in the computations of the statistics by hand described in this chapter, you may want to use data sets that yield integer means and standard deviations. W. P. McGown and Boyd Spencer (1988) as well as Frank Dudek (1988) provide several small data sets that may fill your needs. Bernard Beins (1999) describes a BASIC program he has written for generating data sets with integer means and variances. Finally, John Walsh discusses a FORTRAN program that he uses to create nonnormal data sets when demonstrating how statistical measures can be affected by the shape of the distribution.

Beins, B. C. (1999). A BASIC program for generating means and variances. In M. E. Ware and C. L. Brewer (Eds.), *Handbook for teaching statistics and research methods (2nd ed.)* (pp. 9–10). Mahwah, NJ: Lawrence Erlbaum.

Dudek, F. J. (1988). Data sets having integer means and standard deviations. In M. E. Ware and C. L. Brewer (Eds.), *Handbook for teaching statistics and research methods* (pp. 62–63). Hillsdale, NJ: Lawrence Erlbaum.

McGown, W. P., & Spencer, W. B. (1988). For statistics classes: Data sets with integer means and standard deviations. In M. E. Ware and C. L. Brewer (Eds.), *Handbook for teaching statistics and research methods* (p. 62). Hillsdale, NJ: Lawrence Erlbaum.

Walsh, J. F. (1999). A simple program for generating nonnormal data sets: A FORTRAN program. In M. E. Ware and C. L. Brewer (Eds.), *Handbook for teaching statistics and research methods (2nd ed.)* (pp. 10–12). Mahwah, NJ: Lawrence Erlbaum.

4. To introduce central tendency, Lita Schwartz (1990) asks students what is meant by average? After discussing the terms mean, median, and mode, she has her students explore these central tendency measures by having them keep records of the amount of money they spend on several different budget categories (e.g., food, laundry, movies, recreation) over a four week period. At the end of this period, the students compute the weekly and total means, medians, and modes for each spending category in order to describe their spending habits.

Schwartz, L. L. (1990). Measures of central tendency in daily life. In V. P. Makosky, C. C. Sileo, L. G. Whittemore, C. P. Landry, and M. L. Skutley (Eds.), *Activities handbook for the teaching of psychology* (Vol. 3, pp. 187–188). Washington, DC: American Psychological Association.

5. To introduce descriptive statistics as well as measures of central tendency, Kurt Salzinger (1990) has students contrast the length of words found in novels, textbooks, and/or magazines so that the students can ask questions concerning the manifest difficulty of these various reading materials. Salzinger has students choose a text sample of 200 words from these different sources and then chart the frequency distributions of word lengths as the mean, median, and mode for each text sample. Using this information, students can draw conclusions concerning the reading difficulty (based on word length) for each of these reading materials. The concept of inferential statistics may also be introduced at this point if students want to determine whether the average word length for one source of reading material differs in a meaningful way from another reading source (e.g., *Reader's Digest* versus *New York Times Magazine*).

To demonstrate the real world significance of this activity, Salzinger mentions that this word length measure was used by Mendenhall to resolve authorship disputes between Shakespeare and Bacon. Mendenhall found that the modal length of words used by Shakespeare was four letters while Bacon's words had a mode of three letters. Salzinger points out that Mosteller and Wallace used similar procedures to resolve other authorship disputes.

Salzinger, K. (1990). On the average... In V. P. Makosky, C. C. Sileo, L. G. Whittemore, C. P. Landry, and M. L. Skutley (Eds.), *Activities handbook for the teaching of psychology* (Vol. 3, pp. 185–186). Washington, DC: American Psychological Association.

6. Recognizing that understanding the concept of statistical variability may be a stumbling block for students, Trice, Trice, and Ogden (1990) have developed a series of highly concrete exercises to introduce measures of spread. The first exercise involves having students throw darts at a paper target the size of a legal pad page 15 feet away. A straight line is drawn down the center of this target and students are instructed to aim as close to the center as possible. After seven throws, the paper is removed and the holes in the paper are treated as data points. Students measure how far their throws were from the center line, calculating the mean and variance for these data points. After practicing throwing the darts for fifteen minutes, students repeat the procedure, again measuring how far their throws deviated from the center line target. Trice et al. point out that while the means for both sets of data tend to be similar, the variance does not. Typically, the variance for the second data set is smaller. This observation emphasizes to students the point that both a measure of central tendency as well as a measure of spread is necessary in order to adequately represent a data set.

Another way to introduce the concept of variability in a concrete manner, Trice et al. have their students keep careful track of the amount of time they spend studying each day. After a week of keeping this log, students calculate the mean study time as well the standard deviation associated with this mean study time. You may also want to introduce the concept of standardized scores at this time so that students can compare their mean study time with other students in the class.

Trice, A. D., Trice, O. A., & Ogden, E. P. (1990). Teaching the concept of statistical variability. In V. P. Makosky, C. C. Sileo, L. G. Whittemore, C. P. Landry, and M. L. Skutley (Eds.),

Activities handbook for the teaching of psychology (Vol. 3, pp. 189–191). Washington, DC: American Psychological Association.

7. Kenneth Weaver (1999) uses medical growth charts (which are usually available from pediatricians and local health departments, as well at the web site for the Center for Disease Control (http://www.cdc.gov)) to introduce the concept of variability as well as the normal distribution and standard scores. Students are first asked to estimate the range of weights for healthy newborns and healthy 36-month-old children. Then, using the growth charts, the students compare their estimates of the range and the range between the 5th and 95th percentiles for both newborns and 36-month-old children. Students can see that while the absolute range of weight for the 36-month-olds is greater, the weights of newborns are much more variable. Weaver next uses the growth scale to show how percentiles, *z* scores, and spread are interrelated. For example, students asked why the range in weight between the 10th and 25th percentiles and the 25th and 50th percentiles are the same despite the fact that the first percentile range includes only 15% of the population whereas the second percentile range includes 25% of the population. Weaver also has his students use the charts to convert their birth weights into percentiles and then estimate what the corresponding *z* score would be for their birth weight (estimating both sign and magnitude). Weaver has his students then check on the accuracy of their estimating by converting their percentile into a *z* score by using a table similar to Table B.1 (*z* Values and Their Associated One-Tailed *p* Values) found in Appendix B of the text.

Weaver, K. A. (1999). The statistically marvelous medical growth chart: A tool for teaching variability. *Teaching of Psychology, 26,* 284–286.

MULTIPLE-CHOICE QUESTIONS

1. Cognitive psychologist Stephen M. Kosslyn wrote about three interrelated principles describing how the brain perceives and processes visual information. These three principles can be described as which of the following?
   a. 1. to perceive is to immobilize; 2. mind over matter; and 3. the mind judges a book by its cover.
   b. 1. the spirit is willing, but the mind is weak; 2. mind over matter; and 3. those who exert the first influence upon the mind, have the greatest power
   * c. 1. the mind is not a camera; 2. the mind judges a book by its cover; and 3. the spirit is willing, but the mind is weak.
   d. 1. mind over matter; 2. to perceive is to immobilize; and 3. the mind is not a camera. (185)

2. To visualize the overall pattern of the data, it is often useful to create a
   * a. frequency distribution.
   b. measure of location.
   c. measure of spread.
   d. standardized score. (185-186)

3. A professor is interested in the overall pattern of the scores from an exam he recently gave. Which of the following would be most useful to him?
   a. the mean exam score
   * b. the frequency distribution of the exam scores
   c. the range of the exam scores
   d. the variance of the exam scores (185-186)

4. Eli wants to visually display how student absences fluctuate over the course of the semester. Which type of visual display should Eli use?
   a. stem-and-leaf chart
   b. bar graph
   c. frequency distribution
   * d. line graph (186)

5. A visual display that preserves the original numbers while still providing a visual summary of the data is the
   a. frequency polygon.
   * b. stem-and-leaf chart.
   c. histogram.
   d. frequency distribution. (187)

6. A stem-and-leaf chart showing the data points from two different sets of data is called a _____ stem-and-leaf.
   * a. back-to-back
   b. normal
   c. frequency
   d. comparison (188)

7. A point in a data set which indicates that a known proportion of the data falls above that point and a known proportion of the data falls below that point is called a

a. mean.
* b. percentile.
c. histogram.
d. mode. (189)

8. The measure of central tendency that is determined by identifying the midmost score in a data set is the
a. mean.
b. mode.
c. percentile.
* d. median. (189)

9. Chris calculates the distance between the 25^{th} and 75^{th} percentile in a data set. Chris is interested in identifying the
* a. interquartile range.
b. standard deviation.
c. crude range.
d. extended range. (189)

10. The score that occurs with the greatest frequency in a data set is the
* a. mode.
b. median.
c. mean.
d. variance. (191)

11. In the data set consisting of 3, 3, 3, 5, 5, 9, what is the mode?
* a. 3
b. 4
c. 4.67
d. 6 (191)

12. Data sets that contain two scores that occur with the highest frequency are considered to be
a. symmetrical.
b. asymmetrical.
* c. bimodal.
d. normal. (191)

13. Which of the following is NOT a measure of central tendency?
a. mode
* b. variance
c. median
d. mean (193)

14. In order to describe the typical score, Shellie sums all of the scores and divides that number by the number of scores. Which measure of central tendency has Shellie calculated?
a. mode
b. median
* c. arithmetic mean
d. mean square (191)

15. Which measure of central tendency is most affected by outliers?
    a. range
    b. median
    c. mode
  * d. mean (192)

16. As he calculates the measures of central tendency for his data set, Andrew notices the mean is substantially higher than the median. Andrew concludes that the distribution is most likely
    a. symmetrical.
  * b. positively skewed.
    c. negatively skewed.
    d. bimodal. (192)

17. Using a trimmed mean to measure central tendency may be appropriate when the distribution of the scores is
    a. symmetrical.
    b. normal.
    c. standardized.
  * d. asymmetrical. (192)

18. Which of the following is NOT a measure of spread?
    a. crude range
    b. standard deviation
  * c. standardized range
    d. interquartile range (193)

19. Which measure of spread represents simply the difference between the highest score and the lowest score?
    a. standardized range
  * b. crude range
    c. extended range
    d. interquartile range (193)

20. Which of the following measures of spread is intended to take into account the possible imprecision of the measurement instrument?
    a. the crude range
    b. the interquartile range
    c. the variance
  * d. the extended range (193)

21. Which of the following measures of spread utilizes every score in the data set?
  * a. the variance
    b. the crude range
    c. the interquartile range
    d. the extended range (193)

22. Another name for the variance is the
    a. standard deviation.
  * b. mean square.
    c. extended range.
    d. average deviation. (193-194)

23. Judy subtracts each score from the mean, squares each value, and then determines the mean of these squared values. Which measure of spread is Judy calculating?
    a. the interquartile range
    b. the crude range
    c. the standard deviation
  * d. the variance (193)

24. Which popular measure of spread is calculated by taking the square root of the variance?
    a. the average deviation
    b. the mean square
  * c. the standard deviation
    d. the standard score (194)

25. The root mean square is another name for the
  * a. standard deviation.
    b. variance.
    c. crude range.
    d. interquartile range. (194)

26. If one is interested in measuring the variability in a complete population, one is using a(n) _____ measure. If one is interested in measuring the variability of a sample to generalize to an unknown population, one is using a(n) _____ measure.
    a. standardized; inferential
    b. descriptive; standardized
    c. inferential; descriptive
  * d. descriptive; inferential (195)

27. Which of the following formulae is used in the computation of the confidence interval for the mean?

a. $\sqrt{\frac{(X-M)^2}{N}}$

b. $\frac{X-M}{\sigma}$

* c. $M \pm \frac{(t_{.05})(S)}{\sqrt{N}}$

d. $\frac{\sum (X-M)^2}{N}$ (197)

28. A distribution that can be characterized as a bell shape curve is considered to be a _____ distribution.
a. asymmetric
b. skewed
* c. normal
d. standardized (197)

29. In a normal distribution, the mean is _____ the median.
a. more meaningful than
* b. equal to
c. smaller than
d. larger than (198)

30. In a normal distribution of population scores, roughly how many of these scores will fall between -1σ and +1σ?
a. 2%
b. 50%
* c. 68.3%
d. 95% (198)

31. A standard normal distribution has a mean of _____ and a standard deviation of _____.
* a. 0; 1
b. 1; 0
c. 1; 1.5
d. 0; .5 (198)

32. Steve transforms a score by subtracting the mean from that score and then dividing the result by the standard deviation of all of the scores. The score that Steve has calculated is called the
a. mean.
* b. *z* score.
c. average deviation score.
d. variance. (198)

33. *z* scores statistically translate any individual raw score into a score that indicates how many _____ the score is away from its _____.
* a. standard deviation units; mean
b. variance units; mean
c. standard deviation units; median

d. interquartile units; median (198)

34. The sign of a *z* score indicates whether the individual raw score is larger or smaller than its
a. mode.
b. median.
* c. mean.
d. standard deviation. (198)

35. Scores from data sets with different ranges can be compared and combined using their respective
a. means.
b. variances.
c. standard deviations.
* d. *z* scores. (198)

36. A _____ *z* score is above the mean, while a _____ *z* score is below the mean.
a. negative; positive
* b. positive; negative
c. positive; zero
d. zero; negative (198)

37. An instructor administers an exam to a psychology class. The mean exam grade was 80, with a standard deviation of 10. Supposing you receive a score of 95, what is your *z* score?
a. -1.5
b. -.5
c. +.5
* d. +1.5 (198)

38. An index for reporting the proportion of *N* scores falling on the mode might be as a subscript (in parentheses) of the mode. The name for this index is
* a. modal representativeness index.
b. infinite.
c. precision index.
d. abscissa index. (191)

39. Using an index for reporting the proportion of *N* scores falling on the mode, a mode having 10 scores of "4" out of 50 possible scores falling on the mode would be annotated as
* a. $4_{(.20)}$.
b. $4_{(50).}$
c. μ.
d. $10_{(4)}$. (191)

SHORT ESSAY QUESTIONS

1. Describe the three essential criteria for ensuring graphical integrity when visually displaying one's data.

2. When Steven M. Kosslyn asserts that the mind is not a camera; the mind judges a book by its cover; and the spirit is willing, but the mind is weak, what do these principles mean for researchers when they are presenting data in visual form?

3. What is the benefit of looking at the overall pattern of the data? Describe one method for visualizing the data.

4. Why is it useful to identify the central tendency of a data set? Describe three measures of central tendency.

5. What are outliers? What problems can they create when calculating the arithmetic mean? How might this problem be resolved?

6. Differentiate between symmetrical, positively skewed, and negatively skewed distributions.

7. Why is it important to report the spread of the data along with the central tendency when describing a data set?

8. Differentiate between the crude range, the extended range, and the interquartile range.

9. What does the variance represent? How is this related to the standard deviation?

10. When is it appropriate to use the inferential formula when calculating the variance rather than the descriptive formula?

11. Why would one calculate the confidence interval for a mean?

12. What is a normal distribution? What value does it have to researchers?

13. What is a *z* score? What is the benefit of converting raw scores from different data sets into *z* scores?

CHAPTER 11: *CORRELATING VARIABLES*

CHAPTER OUTLINE

I. What Are Different Forms of Correlations?
- A. The **correlation coefficient** is a single number that can be used to indicate the strength of linear association between two variables (*X* and *Y*).
- B. The **Pearson *r*** is the correlation coefficient frequently used to assess the strength of the linear relationship between two variables.
  1. The Pearson *r* value indicates the strength of the linear relationship.
     - a. A value of 1.0 (positive or negative) indicates that there is a perfect linear relationship between *X* and *Y*.
     - b. A value of 0 indicates that neither *X* nor *Y* can be predicted from the other by using a linear equation.
  2. The sign of the Pearson *r* indicates the direction of the linear relationship.
     - a. A positive *r* value means that increases in *X* are associated with increases in *Y*.
     - b. A negative *r* value means that increases in *X* are associated with decreases in *Y*.
- C. How a correlation coefficient is computed depends on the characteristics of the raw data, such as whether *X* and *Y* are **continuous** or **dichotomous** (or **discrete**) **variables**.
  1. Pearson *r*: two continuous variables
  2. Spearman rho (r_s): two ranked variables
  3. Point-biserial *r* (r_{pb}): one continuous and one dichotomous variable
  4. Phi coefficient (ϕ): two dichotomous variables

II. How Are Correlation Visualized in Scatter Plots?
- A. It is often informative to visualize the degree of linearity between *X* and *Y* by constructing a **scatter plot** (or **scatter diagram**).
- B. If the "cloud" of dots slopes upward, the correlation is positive; if downward, it is negative.
- C. The more tightly clustered the dots are, the greater the correlation between the variables.

III. How Is a Product-Moment Correlation Calculated?
- A. While there are many useful formulas for computing the product-moment correlation coefficient (r), there is one formula using *z* scores that can be quite general and that provides a conceptual definition of the Pearson *r*.
- B. The formula for calculating the Pearson *r* based on *z* scores is:

$$r_{xy} = \frac{\sum z_x z_y}{N}$$

- C. This formula reflects why the Pearson *r* is also called the **product-moment correlation**. The *z*'s in the numerator of the formula are the distances from the means (i.e., "moments") which are multiplied to form "products."

D. Another formula for computing Pearson *r* based on raw scores is:

$$r_{xy} = \frac{N\sum XY - \sum X \sum Y}{\sqrt{\left[N\sum X^2 - (\sum X)^2\right]\left[N\sum Y^2 - (\sum Y)^2\right]}}$$

IV. How Is Dummy Coding Used in Correlation?

A. The **point-biserial correlation** (r_{pb}) is another case of the product-moment *r* (or Pearson *r*) that is computed on data where one variable is continuous while the other variable is dichotomous.

B. The quantification of the two levels of a dichotomous variable is called **dummy coding**.

1. Dummy coding, where numerical values such as 0 and 1 are used to indicate the two distinct parts, is a tremendously useful method because it allows us to quantify any variable that can be represented as dichotomous.
2. Another term for dichotomous is **binary**.

C. The formula for the point-biserial correlation is identical to the one used to calculate the Pearson *r*.

V. When Is the Phi Coefficient Used?

A. The **phi coefficient** is the product-moment correlation used to measure the linear relationship between two dichotomous variables.

B. While the *z* score formula used for the product-moment correlation (Pearson *r*) can be used to calculate the phi coefficient, there is an alternative formula that takes advantage of the fact that the data can be represented as a 2 × 2 table of frequencies (or **counts**) where the cells are labeled A, B, C, and D. This formula is:

$$\phi = \frac{BC - AD}{\sqrt{(A+B)(C+D)(A+C)(B+D)}}$$

VI. How Is a Correlation Calculated on Ranks?

A. **Spearman rho** (r_s) is the correlation coefficient computed when the data are in the form of ranks rather than scores on a rating scale.

B. The formula for computing the correlation coefficient for scores that have been ranked is:

$$r_s = 1 - \frac{6(\sum D^2)}{N^3 - N}$$

Where 6 is a constant value, and *D* is the difference between the ranks assigned to the two scores representing each of the *N* sampling units.

C. If there are no ties in the rankings, $r_s = r$.

LECTURE IDEAS AND ACTIVITIES

1. There are several useful sources for discussing the use (and misuse) of correlations. As mentioned in the text, John Allen Paulos's *Innumeracy* discusses how and why correlation does not imply causation and provides several humorous examples to make his point. Keith Stanovich (2001) devotes a chapter to discussing the correlation and causation issue, providing two illustrations of how making causal inferences based on correlational evidence can have important consequences. For example, Stanovich discusses how educators misinterpreted the directionality of the correlation between eye-movement patterns and reading ability. Based on the correlational evidence, educators assumed that erratic eye-movement patterns caused poor reading and developed special "eye-movement training" programs to enhance children's reading ability. Unfortunately, as Stanovich points out, careful research revealed that the causal relationship between these two variables was in the opposite direction. Stanovich also discusses Joseph Goldberger's investigation of a disease known as pellagra and how he dramatically demonstrated that presumed causal relationship between poor sanitary conditions and this disease was actually due to a third, unrecognized variable—diet.

Paulos, J. A. (1989). *Innumeracy: Mathematical illiteracy and its consequences.* New York: Vintage Books.

Stanovich, K. (2001). *How to think straight about psychology (6th ed.)*. Boston, MA: Allyn and Bacon.

2. Another source of humorous examples to demonstrate that correlation does not imply causality is "Hippocrates' Good News Survey." One of the findings of this survey was that those who did not regularly eat Frosted Flakes as a child were twice as likely to develop cancer as those who regularly did eat Frosted Flakes. In addition, the cancer rate among adults who regularly ate oatmeal as a child was four times greater than among children who did not. These "causal" findings can spark a lively discussion as students try to explain what the real causal factor was in these observed relationships (HINT: age, because cancer tends to develop later in one's life).

Tierney, J. (1987, September/October). Good news! Better health linked to sin, sloth. *Hippocrates*, pp. 30–35.

MULTIPLE-CHOICE QUESTIONS

1. The product-moment (Pearson *r*) correlation coefficient is used to measure the _____ relationship between two variables.
   a. causal
   b. positive
   c. directional
   * d. linear (204)

2. With _____ variables, one can imagine that there can always be another value between two adjacent scores. On the other hand, _____ variables have their values divided into two mutually exclusive categories.
   a. random; dichotomous
   * b. continuous; dichotomous
   c. dichotomous; continuous
   d. continuous; random (204)

3. The _____ of the correlation coefficient indicates the direction of the relationship; the _____ indicates the strength of the relationship.
   a. value; sign
   b. linearity; magnitude
   * c. sign; value
   d. magnitude; linearity (204)

4. Tom wants to determine whether spending more time studying will improve his overall class percentage grade. Statistically speaking, Tom wants to know if there is a _____ between studying and class performance.
   * a. positive correlation
   b. negative correlation
   c. zero correlation
   d. nonlinear correlation (204)

5. Charles finds that the correlation coefficients between his predictor variable and three different outcome variables are +.23, +.45, and –.53. Which correlation coefficient indicates the strongest linear relation that Charles found between his predictor and outcome variables?
   a. +.23
   b. +.45
   * c. –.53
   d. not enough information is provided (204)

6. What is another term for dichotomous?
   * a. binary
   b. correlation
   c. Spearman rho
   d. continuous (208)

7. In a recent study, Mary found a correlation coefficient of .83 between the two variables, foot size and score on the math final. She concludes that these two variables have a strong _____ correlation.
   a. negative
   * b. positive
   c. binary
   d. continuous (204)

8. A _____ correlation coefficient means that decreases in *X* are associated with decreases in *Y*; a _____ correlation coefficient means that decreases in *X* are associated with increases in *Y*.
   * a. positive; negative
   b. zero; negative
   c. negative; positive
   d. positive; zero (204)

9. Donna is calculating the correlation coefficient between two continuous variables. Which correlation coefficient should Donna use?
   * a. Product-moment correlation coefficient
   b. Spearman rho
   c. point-biserial correlation
   d. phi coefficient (205)

10. Steve wants to examine the linear relation between the number of times a commercial appears on television and the dollar sales of the advertised product. Which of the following correlation coefficients should Steve calculate?
   * a. product-moment correlation coefficient
   b. Spearman rho
   c. point-biserial correlation
   d. phi coefficient (205)

11. A graph that can be used to visualize the correlation coefficient is the
   a. line graph.
   b. bar graph.
   c. stem-and-leaf chart.
   * d. scatter plot. (206)

12. In a scatter diagram, the more tightly clustered the data points are around a straight line, the _____ the correlation coefficient is between the two variables.
   a. lower
   * b. higher
   c. closer to zero
   d. weaker (206)

13. Another name for the product-moment correlation is the
    a. Spearman rho.
    b. point-biserial correlation.
    c. phi coefficient.
  * d. Pearson *r*. (204)

14. The correlation computed on two data sets that are in the form of ranks is usually called the
    a. point-biserial correlation.
    b. phi coefficient.
  * c. Spearman rho.
    d. product-moment correlation. (205)

15. Cameron wishes to correlate college football team rankings at the end of this year's football season with the rankings from the previous year. Which correlation coefficient should Cameron use?
  * a. Spearman rho
    b. phi coefficient
    c. point-biserial correlation
    d. any of the above (205)

16. Amala wants to observe the relation between where her teammates placed in one cross-country meet and where they placed in the next cross-country meet against the same opponents. Which correlation coefficient should Amala use to examine this relation?
    a. phi coefficient
    b. point-biserial correlation
    c. product-moment correlation coefficient
  * d. Spearman rho (205)

17. If one wanted to assess the magnitude of the relationship between a continuous variable and a dichotomous variable, one would calculate a
  * a. point-biserial correlation.
    b. Spearman rho.
    c. phi coefficient.
    d. Spearman-Brown coefficient. (205, 208)

18. Brian is interested in whether there is a relationship between one's gender and the length of time spent talking on the phone. Which correlation coefficient should Brian use?
    a. phi coefficient
  * b. point-biserial correlation
    c. Spearman rho
    d. Spearman-Brown coefficient (205, 208)

19. Dummy coding is often used to quantify _____ data.
    a. incomplete
    b. continuous
    c. sample
  * d. dichotomous (208)

20. The correlation coefficient computed between two dichotomous variables is the
    a. Spearman rho.
  * b. phi coefficient.
    c. point-biserial correlation.
    d. Spearman-Brown coefficient. (205, 210)

21. Terri is intrigued about a possible relation between whether one is left-handed or right-handed and one's gender. Which correlation coefficient should Terri calculate to examine this possible relation?
  * a. phi coefficient
    b. point-biserial correlation
    c. Spearman rho
    d. product-moment correlation coefficient (205, 210)

22. A 2 × 2 table of counts can be useful when computing which correlation coefficient?
    a. Spearman rho
    b. point-biserial
  * c. phi coefficient
    d. Spearman-Brown coefficient (211)

SHORT ESSAY QUESTIONS

1. What does it mean to say that two variables are linearly related? How can this relationship be measured?

2. What is the difference between a positive, a negative, and a zero correlation?

3. Why is the Pearson *r* also referred to as the product-moment correlation coefficient?

4. What is dummy coding? Why is it sometimes used to compute a correlation coefficient?

5. What are the four types of correlation coefficients discussed in the text? When is it appropriate to use each one?

6. Why can the Spearman rho, the point-biserial correlation, and the phi coefficient all be thought of as simply variations of the Pearson *r* correlation coefficient?

CHAPTER 12: *UNDERSTANDING p VALUES AND EFFECT SIZE INDICATORS*

CHAPTER OUTLINE

I. Why Is It Important to Focus Not Just on Statistical Significance?
- A. Besides describing data and looking for mutual relations between variables, scientists are often interested in making comparisons using statistical tests.
- B. **Null hypothesis significance testing (NHST)** is the procedure used to decide whether an observed difference between the research groups is due to chance or the result of a "real" difference between the groups.
- C. There has been a growing realization that NHST is sometimes misunderstood or misused.
  1. Finding that an obtained effect is statistically significant at some specified *p* level does not automatically reveal that the effect was sizable or important.
  2. The failure to obtain statistical significance at the desired *p* level does not indicate that there was no obtained effect, or that the obtained effect was trivial or unimportant.
  3. The *p* value of a significance test is influenced not only by the size of the effect, but by the total number of units or observations (*N*).
- D. **Power analysis** is the determination in advance of how many units (e.g., subjects) will be needed to achieve the desired *p* level, given an expected effect size. A similar analysis is often done after the study is completed, in order to estimate its **effective power** (i.e., its actual power).
- E. It is important that attention is not only focused on the *p* value, but also on other aspects of the results, such as the strength of the relation between *X* and *Y*. This relationship is one way of operationally defining the **effect size**.
- F. The text emphasizes the **effect size correlation** ($r_{\text{effect size}}$) as the measure of effect size.
  1. The effect size *r* is easy to calculate from *t*, *F*, and chi-square statistics that meet certain assumptions.
  2. The effect size *r* can be used in situations in which other popular effect size indices may not make sense.
  3. Effect size *r* lends itself to interpretation by a procedure known as the binomial effect size display (or BESD).

II. What Is the Reasoning Behind Null Hypothesis Significance Testing?
- A. Experimenters employing NHST are usually interested in testing the **null hypothesis** (symbolized as H_0) against an **alternative hypothesis** (symbolized as H_1). These two hypotheses are mutually exclusive.
- B. Given the typical experimental design, the null hypothesis would imply no difference in performance between the experimental and control groups. The alternative hypothesis typically is that there is some difference between the experimental and control groups.

III. What Is the Distinction Between Type I Error and Type II Error?

A. **Type I error** implies that one has mistakenly rejected the null hypothesis (H_0) when it is, in fact, true and should not have been rejected.

B. A **Type II error** implies that one has mistakenly failed to reject the null hypothesis when it is, in fact, false and should have been rejected.

C. The risk (or probability) of making a Type I error is called by three different names: **alpha** (α), the **significance level**, and the ***p* value**.

D. The risk (or probability) of making a Type II error is known as **beta** (β).

E. Risks of Gullibility and Blindness

1. Many scientists who do NHST tend to attach greater psychological importance to the risk of making a Type I error than to the risk of making a Type II error.
2. A Type I error is an **error of gullibility**; a Type II error is an **error of blindness**.

IV. What Are One-Tailed and Two-Tailed *p* Values?

A. The text uses the different *p* levels associated with different values of *r* (see Table 12.3 and Table B.5 in Appendix B) to demonstrate how one determines the *p* value.

B. Reported *p* values can be either **two-tailed** or **one-tailed**.

1. A two-tailed *p* value implies that the alternative hypothesis (H_1) did not specifically predict in which side (or tail) of the probability distribution the significance would be detected.
2. A one-tailed *p* value (obtained by halving the two-tailed *p* values) implies that the alternative hypothesis requires the significance to be in one tail rather than in the other tail.

C. There are several options when reporting *p* values.

1. A typically used alternative is to state only that "$p < .05$, two-tail." The problem with this alternative is that it is imprecise.
2. Another alternative is to state that, for example, ".01 < two-tailed $p < .02$."
3. A third alternative, recommended by the *Publication Manual of the American Psychological Association* (2001), is to state the exact *p* value. Scientific notation can be used if *p* is very small.

D. By examining Table 12.3, one can easily discern that a correlation can be significant at $p = .05$, no matter whether it is a very large correlation or a very small correlation.

1. What matters the most is whether the "$N - 2$" is sufficiently large to allow one to detect the particular magnitude of *r* at the desired level of significance.
2. Hence, only reporting that an effect size *r* was "significant" is not completely informative.

V. What Is the Counternull Statistic?

A. The counternull statistic tells us the nonnull magnitude of the effect size that is supported by exactly the same amount of evidence as is the null value of the effect size.

1. Useful for minimizing two common errors in thinking about effect sizes (different from Type I and Type II errors, but are related to them).
    a. Error when a researcher mistakenly infers that a failure to reject the null hypothesis also implies an effect size of zero.

b. The second common error occurs when a researcher mistakenly equates the rejection of the null hypothesis with having demonstrated a scientifically important effect.

B. To compute the CounterNull Statistic:

a. Suppose an experimenter calculated an obtained effect size *r* of .10, with the null hypothesis defined as *r* 5 0, and found *p* 5 .20 (two-tailed). The researcher can use the following formula to estimate the counternull value of a point-biserial *r* (Rosenthal et al., 2000):

$$r_{\text{counternull}} = \sqrt{\frac{4r^2}{1 + 3r^2}}$$

where *r* in the formula is the obtained value of the effect size. Squaring *r* 5 .10 gives us *r*2 5 .01, and therefore

$$r_{\text{counternull}} = \sqrt{\frac{4(.01)}{1 + 3(.01)}} = \sqrt{\frac{.04}{1.03}} = \sqrt{.0388} = .197,$$

which, rounded to .20, is the counternull value that is as likely as the null value of the effect size *r* of zero. Rather than conclude that "nothing happened" because the obtained *p* value exceeded .05, the experimenter instead accepts the conclusion that an effect size *r* of .20 is just as tenable as an effect size of zero. In fact, concluding that the population *r* is closer to .18 would be more defensible than concluding that the population *r* is no different from zero.

To calculate the percentage coverage of the null-counternull interval, we use

$$\% \text{ Coverage} = 100(1.00 - p_{\text{two-tailed}}),$$

which, given an *r* of .10 and an associated two-tailed *p* value of .20, yields 100(1.00 – .20) 5 80%. Had the reported *p* been one-tailed, we would multiply the *p* value by 2 before subtracting it from 1.00, that is,

$$\% \text{ Coverage} = 100[1.00 - 2(p_{\text{one-tailed}})].$$

Our interpretation is that, with 80% confidence, the population value of *r* falls between zero (the null value) and .20 (the rounded counternull value).

VI. What Is the Purpose of Doing a Power Analysis?

A. **Statistical power** has to do with the sensitivity of a statistical test in providing an adequate opportunity to reject the null hypothesis if it warrants rejection.

B. Through a power analysis, one can learn whether there was a reasonable chance of rejecting the null hypothesis, and whether the power should be increased in future studies to increase the sensitivity of the significance test.

C. **Power** represents the probability of not making a Type II error (i.e., 1 – β).

D. For any given statistical test of a null hypothesis, the power of the statistical test is determined by three components:
1. The level of risk of drawing a spuriously positive conclusion (i.e., the *p* level).
2. The size of the study (i.e., the sample size).
3. The effect size.

VII. How Do I Estimate a Confidence Interval for an Effect Size Correlation?

A. The four steps to calculate the CI:
1. Step 1: The $r_{\text{effect size}}$ is transformed to a Fisher z_r (a log-based transformation of *r*).
2. Step 2: Substitute the value of *N* in the study into the following expression:

$$1.96\left(\frac{1}{\sqrt{N-3}}\right)$$

a. The value 1.96 is the standard score *z* for *p* = .05 two-tailed.
b. The other value defines the standard error of Fisher z_r.
3. Step 3: Find the lower limits of the 95% CI by subtracting the result in Step 2 from the Fisher z_r transformed effect size in Step 1. The upper limit is determined by adding the result in Step 2 to the Fisher z_r in Step 1.
4. Step 4: The upper and lower z_r values from Step 3 are transformed back to $r_{\text{effect size}}$ to define the 95% CI around the effect.

B. Smaller *N*s widen the CI; larger *N*s narrow the CI.

VIII. What Can Effect Sizes Tell Us of Practical Importance?

A. Chapter 12 the focus is on certain effect size indicators for use with a 2 x 2 contingency table of independent frequencies.

B. Five effect size indicators that are often reported
1. The *odds ratio* (OR)

$$\text{OR} = \frac{\text{A} / \text{B}}{\text{C} / \text{D}}$$

2. The *relative risk* (RR)

$$\text{RR} = \frac{\text{A} / (\text{A} + \text{B})}{\text{C} / (\text{C} + \text{D})}.$$

3. The *relative risk reduction* (RRR)

$$\text{RRR} = \left[\frac{\text{RD}}{\text{C} / (\text{C} + \text{D})}\right] \times 100,$$

4. The *risk difference* (RD, often described as the *absolute risk reduction*, ARR)

$$\text{RD} = \left(\frac{\text{A}}{\text{A} + \text{B}}\right) - \left(\frac{\text{C}}{\text{C} + \text{D}}\right).$$

5. The *number needed to treat* (NNT), computed as follows: NNT = 1/RD.

C. Note that the effect size *r* appears to be sensitive to both the magnitude of the treatment effect *and* the overall event rate.

1. *Phi*, as described in the previous chapter, is the product-moment correlation, where both variables are scored dichotomously. As illustrated earlier, this *r*-type effect size indicator can be computed directly on the frequencies in a 2 x 2 table of counts by

$$\phi = \frac{BC - AD}{\sqrt{(A + B)(C + D)(A + C)(B + D)}}.$$

IX. What Does Killeen's p_{rep} Tell Me?

A. p_{rep} is the probability of replicating the same direction of a result, given the same size and procedures as the original study.

B. The formula is

$$p_{rep} = \frac{1}{1 + \left(\frac{p}{1 - p}\right)^{\frac{2}{3}}}$$

LECTURE IDEAS AND ACTIVITIES

1. Students may become frustrated by the fact that behavioral research conclusions are stated in terms of probabilities rather than certainties. As Keith Stanovich (2001) points out, most of science is based on probability and as science progresses, more and more "scientific laws" are stated in probabilistic terms. You may want to discuss with students why this is not a shortcoming of behavioral science or science as a whole. You want to mention that much of our daily behavior is based on probabilistic conclusions rather than on certainty. For example, we often make the decision whether or not to carry an umbrella with us based on a weather forecast of rain. We cannot be certain that it will actually rain (unless it is currently raining outside, of course), so our decision to carry an umbrella is based on probability. Using the reported probability of rain, we could decide that it is not going to rain and therefore there is no need to carry an umbrella (i.e., the null hypothesis). We could decide that it is going to rain and therefore one should carry an umbrella (i.e., the alternative hypothesis). Ask students to discuss the probability of rain a weather forecaster would need to give before they conclude that they should carry an umbrella. You can then draw a rough parallel between this decision and the decisions that scientists make concerning hypotheses. Stanovich provides a further discussion of probabilistic reasoning such as the problem of "person-who" statistics (i.e., the invalidation of a probabilistic relationship based on one contrary example).

Stanovich, K. E. (2001). *How to think straight about psychology (6th ed.)*. Boston: Allyn and Bacon.

2. John Allen Paulos (1988) discusses the logic behind statistical significance testing with a specific focus on the trade-offs necessary when evaluating the potential for Type I and Type II errors. For example, he describes decisions made by the Food and Drug Administration with respect to approving a new drug in terms of Type I and Type II errors. Paulos also couches Pascal's wager concerning the existence of God in terms of making a Type I error (reject God when he does exist) versus a Type II error (accepting God when he does not exist). Paulos makes the point that many of life's decisions can be thought of in terms of the probability of Type I and Type II errors.

Paulos, J. A. (1988). *Innumeracy: Mathematical illiteracy and its consequences*. New York: Vintage Books.

MULTIPLE-CHOICE QUESTIONS

1. The initials "NHST" stand for
   a. Normal HypothesiS Testing.
   b. Null HypothesiS Testing.
   c. Normal Hypothesis Significance Testing.
   * d. Null Hypothesis Significance Testing. (219)

2. One problem with NHST is that it can be misconstrued as an indicator of
   a. statistical significance.
   * b. practical importance.
   c. theoretical significance.
   d. statistical differences. (219-220)

3. The *p* value is not sufficient for indicating the importance of the outcome of a study. The reason for this is because
   a. *p* values are too small.
   * b. the size of the study affects the *p* value.
   c. the magnitude of the test statistic must also be known.
   d. effect sizes are related to *p* values. (222-226)

4. The strength of the relationship between *X* and *Y* may be operationally defined as
   a. the *p* value.
   b. power of the study.
   * c. the effect size.
   d. alpha. (220)

5. At the beginning of a study, the researcher hypothesizes that there is no difference between the experimental and control groups. This is an example of a(n)
   a. Type I error.
   * b. null hypothesis.
   c. alternative hypothesis.
   d. significant hypothesis. (221)

6. Statistically speaking, the experimental hypothesis is generally referred to as the _____ hypothesis.
   a. effective
   b. null
   c. working
   * d. alternative (221)

7. Mistakenly rejecting the null hypothesis when it is true and should not have been rejected is technically called a(n)
   * a. Type I error.
   b. Type II error.
   c. alpha error.
   d. beta error. (222)

8. Which of the following is NOT a name for the probability of making a Type I error?

a. alpha
b. significance level
* c. beta
d. *p* value (222)

9. Stacy purchases a new product that claims it will help her sleep better at night. However, after using it, Stacy discovers that it has no effect on her sleeping habits. Which of the following statements would best characterize Stacy's decision to buy the product?
* a. She made a Type I error.
b. She made a Type II error.
c. She made a statistical significance error.
d. She made the correct decision. (222-223)

10. Accepting the null hypothesis when it should have been rejected is an example of
a. a Type I error.
b. alpha.
c. statistical power.
* d. a Type II error. (222)

11. Which of the following refers to the probability of making a Type II error?
a. significance level
* b. beta
c. alpha
d. p-value (222)

12. The "error of blindness" describes which hypothesis decision outcome?
a. Rejecting the null hypothesis when the null hypothesis should not be rejected.
b. Failing to reject the null hypothesis when the null hypothesis should not be rejected.
c. Rejecting the null hypothesis when the null hypothesis should be rejected.
* d. Failing to reject the null hypothesis when the null hypothesis should be rejected. (223)

13. Type I error is to _____ as Type II error is to _____.
a. blindness; gullibility
b. significance; practicality
c. statistical power; effect size
* d. gullibility; blindness (223)

14. When the alternative hypothesis makes a specific prediction, one should use
a. the null *p* value.
b. a two-tailed *p* value.
* c. a one-tailed *p* value.
d. an alternative *p* value. (225)

15. Which of the following examples for reporting *p* values reflects the recommendations of many statisticians?
a. $p \leq .05$
b. $p < .05$
* c. $.01 < p < .05$
d. $p = .032$ (225)

16. _____ tells us the nonnull magnitude of the effect size that is supported by exactly the same amount of evidence as is the null value of the effect size.
    a. The chi-square statistic
    b. The Pearson's r
  * c. The counternull statistic
    d. The coefficient of determination (226-227)

17. The counternull statistic is useful for minimizing which of the following common errors (in thinking about effect sizes)?
    a. the ecological fallacy
    b. inferring that a failure to reject the null hypothesis also implies an effect size of zero
    c. error of equating the rejection of the null hypothesis with having demonstrated a scientifically important effect
  * d. b and c (226)

18. Based on her results, Kathy is unable to reject the null hypothesis, but she does feel that the effect size looks promising. What should Kathy do next?
    a. recalculate the effect size
    b. use a less stringent alpha level
  * c. do a statistical power analysis
    d. reevaluate her hypothesis (227)

19. Power is the probability of not making a(n)
    a. Type I error.
  * b. Type II error.
    c. statistical error.
    d. experimental error. (228)

20. Power is equal to
    a. alpha.
    b. beta.
    c. 1 minus alpha.
  * d. 1 minus beta. (228)

21. The power of a statistical test is determined by all of the following factors EXCEPT
    a. the *p* value associated with the statistical test.
  * b. the internal validity of the study.
    c. the sample size used in the study.
    d. the effect size associated with the statistical test. (228)

22. Through statistical power analysis, one can estimate the _____ needed to adequately test a hypothesis.
  * a. number of participants
    b. effect size
    c. significance test
    d. *p* value (228)

23. Tamara wants to know how many subjects she should use in her study. To answer this question, Tamara should

a. calculate the study's effect size.
b. use null hypothesis significance testing.
* c. conduct a power analysis.
d. construct a BESD. (228)

24. Smaller sample sizes tend to _____ confidence intervals, whereas larger sample sizes tend to _____ confidence intervals.
a. have no effect on; shrink
b. widen; have no effect on
c. shrink; widen
* d. widen; shrink (229-230)

25. Killeen's p_{rep} indicates the probability of replicating the same _____ of effect of the original study, given the same *N* and procedures used.
a. size
b. outcome
* c. direction
d. power (234)

SHORT ESSAY QUESTIONS

1. How is null hypothesis significance testing (NHST) sometimes misunderstood or misused?

2. Why does "nonsignificance" not mean the same thing as "no effect" in null hypothesis significance testing?

3. What role does probability have in significance testing?

4. What is the difference between Type I and Type II errors? Which error do scientists seem to be more concerned with? Why?

5. Why are Type I and Type II errors characterized respectively as being errors of gullibility and blindness to a relationship?

6. When would one use a one-tailed *p* value rather than a two-tailed *p* value? Why would one prefer to use a one-tailed *p* value in these situations?

7. What are the two common errors in thinking about effect sizes that the counternull statistic helps the researcher minimize? Is this the same as Type I and Type II errors?

8. What is meant by the power of a statistical significance test?

9. How can a statistical power analysis be useful when planning an experimental study?

10. Why would one want to calculate the confidence interval associated with an observed effect size?

CHAPTER 13: *THE COMPARISON OF TWO CONDITIONS*

CHAPTER OUTLINE

I. What Do Signal-to-Noise Ratios Have to Do With *t* Tests?

A. The choice of a statistical test is determined by the nature of the research question and the design of the study.

B. The ***t* test** (or **Student's *t***) tests the likelihood that the population means represented by the two groups are equal (i.e., the null hypothesis), by setting up a **signal-to-noise ratio**.

1. The "signal" is represented by the difference between the two means.
2. The "noise" is represented by the variability of the scores within the samples.
3. The larger the signal is relative to the noise, the more likely the null hypothesis is to be rejected.

C. The *t* test examines the difference between the two group means (i.e., the strength of the "signal") against the background of the within-group variability (i.e., the amount of "noise").

1. The larger the difference between the means (i.e., the stronger the signal), or the smaller the within-group variability (i.e., the lower the noise) for any given study, the larger the resulting *t* value will be.
2. The larger the *t* value, the more likely the difference between the means of the two groups will be considered to be "statistically significant" as larger *t* values are associated with a lower level of probability (*p* value or alpha).

II. How Do I Compute an Independent-Sample *t* Test?

A. Two groups are assumed to be **independent** of each other if the results in one group are not influenced by the results in the other group.

B. The general purpose formula for calculating *t* is:

$$t = \frac{M_1 - M_2}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2} S^2}}$$

1. M_1 and M_2 are the means of the two independent groups.
2. n_1 and n_2 are the number of units (the number of participants) in each of the two groups.
3. S^2 is the *unbiased estimator of the population variance* (i.e., a single estimate of the variance associated with both populations from which the two groups were drawn).
4. The formula for computing S^2 is:

$$S^2 = \frac{\sum (X_1 - M_1)^2 + \sum (X_2 - M_2)^2}{n_1 + n_2 - 2}$$

III. What Can a Table of *p* Values for *t* Teach Me?

A. The *t* test can be considered to be a family of curves (or *t* distributions) in addition to being a single test of statistical significance. There is a different curve, each

resembling the standard normal distribution, for every possible value of the **degrees of freedom** (*df*) of the *t* test.

B. William Gosset, inventor of the *t* test, calculated the curve for each number of degrees of freedom associated with *t*. Using a table based on these calculations, one can easily determine the probability of obtaining a particular *t* value with a specified *df* if the population mean difference was truly zero.

C. The *t* distribution gradually approximates the standard normal distribution as the sample size increases.

D. Based on NHST convention, the .05 significance level is used as the demarcation point such that obtained *t* values that are associated with a probability value of .05 or less are considered to be "statistically significant."

IV. What Is an Effect Size Index for an Independent-Sample *t*?

A. Cohen's *d* is estimated by the following formula:

$$\text{Cohen's } d = \frac{M_1 - M_2}{\sigma_{\text{pooled}}}$$

B. and the pooled population s is

$$\sigma_{\text{pooled}} = S_{\text{pooled}} \left(\sqrt{\frac{df}{N}} \right).$$

V. How Do I Interpret Cohen's *d* for Independent Groups?

A. There are convenient formulas for converting *d* into *t* and into an effect size *r*.

B. Cohen suggested that *d* of .2, .5, and .8 be considered small, medium, and large effects, but context determines the practical importance of the effects.

1. It is essential to specify the particular index that you are using, and think twice before using these labels for effect sizes, as they may be misconstrued as implying that "small" means inconsequential. Cohen (1988) cautioned that "the *meaning* of any given ES [effect size] is, in the final analysis, a function of the context in which it is embedded" (p. 535).

VI. How Do I Compute Interval Estimates for Cohen's *d*?

A. For Cohen's *d* on independent means, we obtain the 95% confidence interval (95% CI) for *d* by

$$95\%\ \text{CI} = d \pm t_{(0.5)} (S_{\text{Cohen's } d}),$$

where t(.05) is the critical value of t at p = .05 two-tailed for $df = n_1 + n_2 - 2$, and $S_{\text{Cohen's d}}$ is the square root of the variance of Cohen's *d*, given by

$$S^2_{\text{Cohen's } d} = \left[\frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(df)} \right] \frac{n_1 + n_2}{df}.$$

B. The counternull value of Cohen's *d* is easier to estimate because it is simply twice the obtained *d* in most cases (Rosenthal & Rubin, 1994).

VII. How Can I Maximize the Independent-Sample *t*?

A. The *t* test, like any significance test, can be shown to consist of two components, one having to do with the effect size and the other with the size of the study.

1. In other words, a *t* test can be expressed by the following conceptual equation:

Significance test = Size of effect × Size of study

2. Another formula for the *t* test in which *t* is mathematically broken down into an effect size and a study size component is:

$$t = d \times \left(\frac{\sqrt{n_1 n_2}}{n_1 + n_2} \times \sqrt{df} \right)$$

or, if the two groups are the same size,

$$t = d \times \frac{\sqrt{df}}{2}$$

B. Examination of this alternative equation for calculating *t* suggests three ways one can optimize *t*.

1. Drive the means further apart.
2. Decrease the variability within groups.
    a. The variability of responses can be decreased by standardizing the research procedures in order to make them more uniform.
    b. Variability may also be decreased by recruiting subject samples that are fairly homogeneous in those characteristics that are substantially associated with the dependent variable.
3. Increase the total size of the study.

VIII. How Does a Paired *t* Test Differ From an Independent-Sample *t* Test?

A. When samples are not independent (i.e., there is a degree of prior relatedness between the scores in Group 1 and those in Group 2), the independent *t* test yields a biased value.

B. When comparing the means for samples that are not independent, one uses a **paired *t* test** (also called a **one-sample *t*-test** or a **correlated *t*** or **matched *t***).

C. The formula for the paired *t* test is:

$$t = \frac{M_D}{\sqrt{\left(\frac{1}{N}\right) S_D^2}}$$

1. M_D is the mean of the $D = X_1 - X_2$ scores.
2. N is the number of D scores (i.e., the number of lined-up pairs).
3. S_D^2 is the unbiased estimate of the population value of σ_D^2, with S_D^2 defined by

$$S_D^2 = \frac{\sum (D - M_D)^2}{N - 1}$$

D. To determine the *p* value as well as $r_{\text{effect size}}$ associated with this paired *t* value, one would use the same procedures employed for the independent *t* test. The one

difference is that the degrees of freedom are defined as $N - 1$, where N is the number of paired scores.

IX. What Is an Effect Size Index for a Paired *t*?

A. Just as the independent-sample *t* was shown to consist of a size-of-effect component and a size of-study component, parsing the paired *t* reveals a similar conceptual relationship:

$$t = d \times \sqrt{N - 1},$$

where the size-of-effect component is Cohen's *d* for paired observations, and the size-of-study component is the number (*N*) of paired observations minus 1. Cohen's *d* for paired observations is given by

$$d = \frac{M_D}{\sigma_D}$$

where M_D is again the mean of the $D = X_1 - X_2$ scores, and s_D is the standard deviation of the *D* scores, defined as

$$\sigma_D = \sqrt{\frac{\Sigma(D - M_D)^2}{N}}.$$

If we rearrange the conceptual equation for the paired *t*, it follows that this version of a Cohen's *d* can be obtained from a paired *t* by

$$d = \frac{t}{\sqrt{N - 1}} = \frac{t}{\sqrt{df}}.$$

Using the conceptual equation for a paired *t*, we compute

$$t = d \times \sqrt{N - 1}$$

as a check on our calculations, we compute d from the paired t by

$$\text{Cohen's } d = \frac{t}{\sqrt{N - 1}}$$

to express the effect size in units of r, the same formula illustrated earlier in this chapter for estimating an effect size r from an independent sample t can be used:

$$r_{\text{effect size}} = \sqrt{\frac{t^2}{t_2 + df}}$$

B. Statistically, the interpretation of the effect size *r* based on correlated observations is more complex than the effect size *r* from the independent-sample *t.*

LECTURE IDEAS AND ACTIVITIES

1. To introduce the *t* test, the instructor may want to develop data sets based on variables that are of intrinsic interest to the class (see also the lecture ideas and activities for Ch. 10). For example, the instructor could introduce the *t* test in the context of asking the class to determine whether the males (or females) in the class are significantly taller than the females (or males) or whether one bus route from point A to point B is, on average, quicker than another. Another possibility is to use a data set that is available on the Internet. For example, the Data and Story Library is an online collection of data files from actual studies that can be used to illustrate a wide variety of basic statistical methods. The Internet address for this resource is http://lib.stat.cmu.edu/DASL/.

2. To illustrate the use of the correlated *t* in a situation in which the independent *t* test is clearly inappropriate, the instructor may want to develop pretest and posttest data. The correlated *t* is appropriate for comparing these data, whereas the independent *t* can be useful for comparing the posttest scores in two independent groups. The independent *t* can also be useful for comparing the posttest-minus-pretest scores in the experimental group with the posttest-minus-pretest scores in the control group.

MULTIPLE-CHOICE QUESTIONS

1. Which of the following statistical tests would one use to compare the means from two groups?
   a. Pearson *r*
   * b. *t* test
   c. chi-square
   d. *z* score test (238)

2. The *t* test is used to compare the means from how many groups?
   * a. 2
   b. 3
   c. 4
   d. 5 (238)

3. In a signal-to-noise ratio, the "signal" represents the
   a. number of participants in both groups.
   b. number of participants minus some value.
   c. variability in scores within the groups.
   * d. difference between the means of both groups. (238)

4. The *t* distribution gradually approximates _____ as the sample size increases.
   a. the bifurcated distribution.
   * b. the standard normal distribution.
   c. the chi-square distribution.
   d. both a and c. (243)

5. A waitress is interested in whether she receives larger tips in general from her male customers than her female customers. She records the gender of the customer who paid the bill and the amount of the tip over a two-day period. Statistically speaking, the **"signal"** in this research scenario would be
   a. the difference between the average tip amounts received on the first and second days.
   b. the fluctuation in tip amounts received from male customers during the two-day period.
   * c. the difference between the average tip amounts received from the male and female customers.
   d. the fluctuation in tip amounts received from the male and female customers. (238-239)

6. In a signal-to-noise ratio, the "noise" represents the
   a. difference between the means of both groups.
   b. variability in means between the groups.
   * c. variability in scores within the groups.
   d. number of participants in both groups. (238)

7. A waitress is interested in whether she receives larger tips in general from her male customers than her female customers. She records the gender of the customer who paid the bill and the amount of the tip over a two-day period. Statistically speaking, the **"noise"** in this research scenario would be
   a. the difference between the average tip amounts received on the first and second days.
   b. the fluctuation in tip amounts received from male customers during the two-day period.
   c. the difference between the average tip amounts received from the male and female customers.
   * d. the fluctuation in tip amounts received from the male and female customers. (239)

8. Which of the following conditions will result in larger *t* values?
   a. smaller difference between means/smaller within group variance
   b. smaller difference between means/larger within group variance
   c. larger difference between means/larger within group variance
   * d. larger difference between means/smaller within group variance (240, 251-252)

9. When the results in one group are not influenced by the results in the other group, the samples are considered to be
   a. random.
   * b. independent.
   c. dependent.
   d. separate. (240)

10. Brian wants to compare the test scores in one class to the test scores in another class. Which test of statistical significance should Brian use?
   a. Pearson *r*
   * b. an independent *t* test
   c. a correlated *t* test
   d. a point-biserial correlation (239-240)

11. In which of the following situations would one use an independent *t* test?
   * a. A professor wants to know if one of his two Introduction to Psychology classes scored significantly higher than the other class on the same exam.

b. A professor wants to know if there was a significant improvement in test scores from Exam 1 to Exam 2 for one of his Introduction to Psychology classes.
c. An educator wants to know if students spend more time talking on the phone than studying during a typical day.
d. A researcher wishes to learn whether the oldest child in a family does significantly better on standardized tests than the youngest child in the family. (239-242)

12. Uriel wants to know if the men in his class spend more time watching sports on television than the women. Uriel asks 13 women in his class to keep track of the number of hours they spent watching sports on television during a seven-day period. He asks 12 men in his class to do the same. What would be the degrees of freedom associated with the *t* test in this situation?
a. 13
* b. 23
c. 24
d. 25 (242-243)

13. Harry is using an independent *t* test to compare the means of two groups. There are ten participants in each group. How many degrees of freedom would be associated with this statistical test?
a. 8
b. 9
* c. 18
d. 19 (242-243)

14. Which of the following measures represents the "noise" component in the *t* test?
a. the population variance
b. the population standard deviation
* c. the unbiased estimator of the population variance
d. the unbiased estimator of the population standard deviation (242)

15. With 8 degrees of freedom, Linda finds that the probability associated with her obtained *t* value of 1.9 is $p = .047$ (one-tailed). What should Linda conclude?
a. The means from both groups are not significantly different from each other.
* b. The means from both groups are significantly different from each other.
c. The difference between the means of both groups has practical significance.
d. She did not have enough participants in her study. (243)

16. The *p* value associated with a statistical test indicates _____ while the associated effect size indicates _____.
a. practical importance; statistical significance
b. statistical significance; statistical importance
* c. statistical significance; practical importance
d. practical importance; practical significance (242-244)

17. The effect size computed for the *t* test is equivalent to _____.
    a. the Pearson *r*
    b. the phi coefficient
    c. the Spearman rho
  * d. the point-biserial correlation (247)

18. A professor wants to compare the test scores of his two psychology classes (Class A and Class B). Each class has 20 students. The professor calculates a *t* value of 2.3 and concludes that Class B has a mean test score that is significantly higher than Class A. What is the effect size associated with this statistical test?
    a. .12
    b. .24
  * c. .35
    d. .48 (244-246)

19. Significance tests have two components. They are:
    a. size of effect; independence
    b. independence; dependence
  * c. size of effect; size of study
    d. *p* value; test statistic (250)

20. One can optimize and strengthen the power of a *t* test in all of the following ways, EXCEPT
    a. by using stronger treatments to drive the means of the two groups farther apart.
  * b. by decreasing the effect size of the study.
    c. by decreasing the variability within groups.
    d. by increasing the total size of the study. (244-251)

21. Rather than personally reading the research instructions to each subject, an investigator has the subjects listen to a tape recording of the research instructions. In this case, the researcher is trying to optimize *t* by
    a. driving the means further apart.
  * b. decreasing the variability of response within groups.
    c. using a paired *t* test rather than an independent *t* test.
    d. increasing the total size of the study. (251)

22. In a test of the impact of delay on recall, a researcher has two groups study a list of words. For the immediate recall condition, the researcher has the subjects recall the words immediately after the study period has ended. For the delay recall condition, the researcher has the subjects recall the words three minutes after the study period has ended. The researcher found that there was not a significant difference in the number of words recalled between the two conditions. The researcher decides to conduct the study again, this time using fifteen minutes as the delay between studying and recalling the list. The researcher is trying to optimize *t* by
    a. increasing the total size of the study.
    b. using the dependent *t* test rather than independent *t* test.
    c. decreasing the variability of response within groups.

* d. driving the means further apart. (250-251)

23. Ed wants to know if students did significantly better on a test the second time they took it. Which test of statistical significance should Ed use?
    a. Cohen's *d*
    b. Pearson *r*
    c. an independent *t* test
* d. a paired *t* test (251-252)

24. William has 20 students take a test on two different occasions. What would be the *df* if he wanted to compare the mean scores from both occasions?
    a. 39
    b. 38
* c. 19
    d. 18 (252)

25. Paul wants to know if students in his class spend more time watching basketball or football on television. Paul has 13 women and 12 men record the numbers of hours they spent watching football and basketball during a seven-day period. What would be the *df* associated with the *t* test in this situation?
    a. 13
    b. 23
* c. 24
    d. 25 (252)

SHORT ESSAY QUESTIONS

1. What is the difference between independent and nonindependent samples? Which *t* test would one use under each circumstance?

2. What is meant by a "signal-to-noise ratio"? Why does this characterize statistical significance testing using the *t* test?

3. What does the *p* value associated with an obtained *t* value indicate?

4. Why is it as important to calculate the effect size in addition to evaluating the statistical significance of an obtained *t* value?

5. Explain the conceptual relationship: *Significance test = Size of effect × Size of study.*

6. Describe the three ways that one can strengthen or maximize *t*.

CHAPTER 14: *COMPARISONS OF MORE THAN TWO CONDITIONS*

CHAPTER OUTLINE

I. What Is Analysis of Variance (ANOVA), and How Are *F* and *t* Related?

A. Although the *t* test is often used whenever there are only two means to be compared, this test can be used to examine a predicted trend in more than two conditions.

B. Another very popular statistic is the *F* test. The *F* test is a signal-to-noise ratio that divides up variability in a procedure called **analysis of variance**, or **ANOVA**.

C. *F* and *t* are related in that squaring *t* always produces *F*, although taking the square root of *F* does not always produce *t*. When there are only two groups to be compared, taking the square root of *F* always produces *t*.

D. Because $F = t^2$ whenever there are two groups being compared, the effect size of *F* in these situations can be computed as

$$r_{\text{effect size}} = \sqrt{\frac{F}{F + df}}$$

where *df* refers to the degrees of freedom "within conditions," obtained by determining the *df* within each group (or condition) and then adding them.

E. The analysis of variance provides a formal comparison of the variation between the average results per condition and the average variation within the different conditions.

F. In this kind of analysis, a ratio (the ***F* ratio**, or ***F* test**) is formed of S^2_{between} and S^2_{within}.

G. *F* ratios usually have values close to 1.0 when the variation between conditions is not different from the variation within conditions (i.e., when H_0 is true). The larger the *F* ratio becomes, the greater the dispersion of group means relative to the dispersion of scores within groups.

II. How Is Variability Apportioned in a One-Way ANOVA?

A. One purpose of the ANOVA is to divide up the variation of all the observations into a number of separate sources of variance.

B. *SS* is the abbreviation for **sum of squares**.

C. Total *SS* = between-conditions *SS* + within-conditions *SS*.

1. Total *SS* is the sum of squares of deviations of all the measurements from the grand mean, or

$$\text{Total } SS = \sum (X - M_G)^2$$

where *X* is each observation and M_G is the grand mean (i.e., mean of the condition means).

2. Between-conditions *SS* is the sum of squares of the deviations of the condition means from the grand mean, or

$$\text{Between } SS = \sum \left[n_k (M_k - M_G)^2 \right]$$

where n_k is the number of observations in the *k*th condition (and *k* is *any* particular condition), M_k is the mean of the *k*th condition, and M_G is the grand mean.

3. The within-conditions *SS* is the sum of squares of deviations of the measurements from their condition means, or

$$\text{Within } SS = \sum (X - M_k)^2$$

where X is each observation and M_k is the mean of the condition to which X belongs.

III. How Are ANOVA Summary Tables Set Up and Interpreted?

A. The rows label the source of variation.

B. The sum of squares for each source of variation is listed in the *SS* column.

C. The degrees of freedom (*df*) are listed in the next column.

1. The degrees of freedom between conditions is defined as $df_{between} = k - 1$.
2. The degrees of freedom within conditions is defined as $df_{within} = N - k$.
3. The total degrees of freedom is defined as the total number of measurements minus 1, or *df* total $= N - 1$.

D. The next column, *MS*, shows the mean squares for each source of variation. This column is determined by dividing the sum of squares by the corresponding *df*.

1. The *MS* values can be seen as the amounts of the total variation (measured in *SS*) attributable to each *df*.
2. The larger the *MS* for the between-condition source of variance (the signal) relative to the within-condition source of variance (the noise), the less likely becomes the null hypothesis of no difference between the conditions.

E. The *F* in the next column represents the ratio of the two mean squares.

1. The denominator (i.e., the mean square for error) serves as a kind of base rate for noise level, or typical variation.
2. The numerator (i.e., the signal) is a reflection of both the size of the effect and the size of the study.

F. The final column in the ANOVA summary table provides the probability associated with obtaining the corresponding *F* value of this magnitude or larger if the null hypothesis of no difference was true.

1. For every combination of *df* between and *df* within, there is a different probability curve.
2. *F* is intrinsically one-tailed as a test of significance.

IV. How Can I Test for Simple Effects After an Omnibus *F*?

A. While the *F* test indicates the probability of the group means being that different from each other if the null hypothesis was true, it does not indicate specifically which group or groups differ from the others when there are more than two comparison groups.

B. Comparisons of specific group means are called **tests of simple effects**. An easy way to do these tests is to do *t* tests.

1. The S^2 value in the formula used to calculate *t* is the within-conditions *MS*.
2. Because the within-conditions *MS* represents the pooled estimate of S^2, the *df* associated with the within-conditions *SS* is used as the *df* in the test of significance of *t*.

C. If one had been planning to compute a specific *t* test from the beginning, one can do so, regardless of whether the overall *F* was significant or not.

1. However, if one is exploring for large differences that were not specifically predicted, *t*-test results are much more interpretable if the overall *F* is

significant, as some differences will be statistically significant by chance alone.

2. To protect against chance findings of significant *t* values, researchers will sometimes use a more conservative level of significance than the .05 level (e.g., .01 or .001).
3. Nevertheless, it is recommended that one's statistical decisions are not based on the significance value alone.
    a. One should also consider the effect size and its corresponding confidence interval and BESD.
    b. When calculating the effect size correlation, one uses the formula used for *t* tests, but the degrees of freedom are now defined from the groups being compared.

V. How Is Variability Apportioned in a Two-Way ANOVA?

A. R. A. Fisher, the inventor of the *F* test, observed that it is sometimes possible to rearrange a one-way design to form a two-way design of much greater power to reject certain null hypotheses.

B. There are several benefits associated with using factorial designs.

1. One can answer more questions than is possible with one-way designs. For example, one can learn whether the effects of one factor are similar for each of the two or more conditions of the other factor.
2. When two or more factors are being examined simultaneously, the participants are, in a sense, serving "double duty," resulting in an increase in the power of the statistical tests concerning the **main effects**.
3. Factorial designs allow for the examination of possible **interaction effects**. The interaction, although it represents the combination of the independent variables, is (in a statistical sense) comprised of "leftover effects" called **residuals**.

VI. How Do I Interpret Main and Interaction Effects?

A. The **grand mean** is the mean of all group means.

B. The **row effect** is the mean of that row minus the grand mean.

C. The **column effect** is the mean of that column minus the grand mean.

D. The **interaction effect** is leftover effects after the grand mean, the row effect, and the column effect are subtracted from the group mean.

VII. How Do I Compute a Two-Way ANOVA and Set Up a Summary Table?

A. The total sum of squares is computed as:

$$\text{Total } SS = \sum (X - M_G)^2$$

where X is each observation or measurement and M_G is the mean of all the condition means.

B. The within-conditions *SS* is computed as:

$$\text{Within } SS = \sum (X - M_k)^2$$

where M_k is the mean of the group or condition to which each observation or measurement (X) belongs.

C. The sum of squares of the rows is defined as:

$$\text{Row } SS = \sum[nc(M_r - M_G)^2]$$

where n is the number of observations in each condition; c is the number of columns contributing to the computation of M_r (the mean of the rth row); and M_G is the grand mean.

D. The sum of squares of the column is defined as

$$\text{Column } SS = \sum[nr(M_c - M_G)^2]$$

where n is the number of observations in each condition; r is the number of rows contributing to the computation of M_c (the mean of the cth row); and M_G is the grand mean.

E. The interaction sum of squares is defined as:

$$\text{Interaction } SS = \text{total } SS - (\text{row } SS + \text{column } SS + \text{within } SS)$$

F. Error refers to the magnitude of the deviations between the group mean and the individual scores. That is,

$$\text{Error} = \text{score} - \text{group mean}$$

G. So, based on the additive model,

$$\text{Score} = \text{grand mean} + \text{row effect} + \text{column effect} + \text{interaction effect} + \text{error}$$

H. Computing the degrees of freedom:

1. $df_{rows} = r - 1$, where r is the number of rows.
2. $df_{columns} = c - 1$, where c is the number of columns
3. $df_{interaction} = (r - 1)(c - 1)$
4. $df_{within} = N - k$, where N is the total number of observations or measurements, and k is the number of groups or conditions.

I. Effect size of each F (using formula from beginning of this chapter):

$$r_{\text{effect size}} = \sqrt{\frac{F}{F + df_{within}}}$$

VIII. What Are Contrasts, and How Do I Compute Them on More Than Two Groups?

A. t and F can be used to make focused comparisons (called **contrasts**) of more than two conditions.

B. Contrasts not only address precise predictions but also allow one to compute meaningful effect size measures when comparing more than two groups.

C. For example, one can compute a t or F to test a predicted linear pattern of regularly increasing means among the group conditions.

1. This prediction is stated in simple integers (called contrast weights, or **lambda coefficients**, or **λ weights**) that sum to zero (i.e., $\sum\lambda = 0$).
2. One can then compare the predicted contrast weights with the obtained scores using the following formula:

$$t_{\text{contrast}} = \frac{\sum M\lambda}{\sqrt{MS_{within}(\sum\frac{\lambda^2}{n})}}$$

where M in the numerator refers to a specific condition mean; MS_{within} in the denominator refers to the within-conditions mean square, n is the number of

observations in the condition, and λ refers to the predicted contrast weight for that condition.

3. While squaring the $t_{contrast}$ will produce the $F_{contrast}$, one can compute the $F_{contrast}$ directly from the raw data by first calculating the contrast mean square ($MS_{contrast}$) from

$$MS_{contrast} = \frac{nL^2}{\sum \lambda^2}$$

where

$$L = M_1\lambda_1 + M_2\lambda_2 + M_3\lambda_3 + \ldots M_k\lambda_k$$

and then dividing $MS_{contrast}$ by MS_{within}.

IX. What Do $r_{effect\ size}$, $r_{alerting}$, and $r_{contrast}$ Tell Me?

A. There are several correlational indices that are informative when interpreting the meaning of contrasts when working with more than two groups.

1. The $r_{effect\ size}$ is the correlation between an individual's score (Y) on the dependent variable and the contrast weight (l) assigned to the condition in which the individual belongs. To compute $r_{effect\ size}$ from the contrast F, this index is also denoted as r_{Yl}, can be computed by:

$$r_{effect\ size} = r_{Y\lambda} = \sqrt{\frac{F_{contrast}}{F_{contrast} + F_{noncontrast}(df_{noncontrast}) + df_{within}}},$$

2. The **alerting correlation** ($r_{alerting}$) is the correlation between the means and their respective contrast weights.
3. The **contrast correlation** ($r_{contrast}$) is the partial correlation between scores on the dependent variable and the lambda coefficients after removing any other patterns in the data. In two-group designs, $r_{contrast} = r_{effect\ size}$.

X. How Are Contrasts on Multiple Repeated Measures Computed?

A. In **intrinsically repeated measures research**, participants must be measured two or more times in order to address the question of interest.

B. One can create contrast scores for each subject to test specific predictions. The text provides an example where the prediction is that subjects would improve by an equal amount on each occasion of measurement.

1. The contrast score, or L score, indicates the degree to which that particular subject behaved in accordance with the prediction.
2. The L score is simply the sum of the products of the contrast weights, multiplied by the subject's actual performance, or

$$L = \Sigma\lambda = Y_1\lambda_1 + Y_2\lambda_2 + \ldots Y_k\lambda_k$$

3. One computes a one-sample t test using all of the L scores for the subjects:

$$t_{(df)} = \frac{M_L}{\sqrt{\left(\frac{1}{N}\right)S_L^2}}$$

where M_L is the mean of the L scores, N is the number of subjects, S_L^2 is the variance of the L scores, and *df* refers to the degrees of freedom for the one-sample t test, usually $N - 1$.

C. An alternative to the use of L scores in repeated measures analyses as long as there are at least three occasions of measurements is the use of *r*s. These *r*s are simply the correlations of the repeated measurements with their associated λ weights. The formula is:

$$t_{(df)} = \frac{\bar{r}}{\sqrt{\left(\frac{1}{N}\right)S_r^2}}$$

XI. How Are Latin Square Designs Analyzed?

A. **Nonintrinsically repeated measures** research is repeated measure research in which it was not actually essential to use repeated measures, but their use increases the efficiency, precision, and the statistical power of the study.

B. To separate order effects from condition effects, we can counterbalance the design by the use of the Latin square described in Chapter 7.

1. An example is the Latin square design described at the end of this chapter.

LECTURE IDEAS AND ACTIVITIES

1. As with teaching about the *t* test, it may be easier for students to understand the logic behind the analysis of variance if the students are using data to answer questions of interest to them. For example, sports box scores may be used to test whether there are differences between different divisions in a sports league with respect to team wins (e.g., a comparison of wins between the East, Central, and Western divisions of Major League Baseball's National League).

2. To show the equivalence between t^2 and F, the instructor can display two sets of results. In Results A there are two independent groups, and in Results B there are three independent groups. We can compute an F test on Results A or Results B, but we can compute a t test only on Results A. Thus we can square our t value to obtain the value of the F test in results A, and we can take the square root of the F test in Results A to get the t value. But we cannot take the square root of the F value in Results B to get a t test; it would not make sense because $k > 2$.

3. Mark Sciutto (2000) describes an activity he uses to concretely demonstrate the impact of various factors on the F ratio. This activity is an extension of an activity Johnson (1989) uses to help students intuitively understand the concepts of between- and within-group variance. After discussing the one-way ANOVA, Sciutto shows two lightweight cardboard boxes to his class. One of these boxes is labeled "Between" while the other is labeled "Within." Sciutto then shows his class an assortment of small objects of variable weights. Sciutto informs his class that some of these objects will be used to represent individual differences (he uses small action figures),

some will represent measurement error (he uses stopwatches), and other objects will represent various treatment effects (he uses different size batteries). After the class inspects the objects, a student is asked to come to the front of the class and close his or her eyes. Sciutto places the measurement error and individual difference objects into both boxes and asks the student to decide whether or not the two boxes differ in weight. The student is also asked to indicate how certain he or she is in his or her decision (0 to 100%). Sciutto then proceeds to systematically manipulate the content of the boxes, asking a student volunteer again to compare the relative weight of the two boxes. For example, to represent a small treatment effect, Sciutto places a "AAA" battery into the "between" box. A medium treatment effect is represented with a "C" battery while a very large effect is represented with two "D" batteries. Sciutto has his students again make these comparisons after removing the individual differences and measurement error objects from both boxes. Scuitto has found that this simple concrete demonstration has helped his students better understand concepts such as effect size, power, and errors in hypothesis testing.

Johnson, D. E. (1989). An intuitive approach to teaching analysis of variance. *Teaching of Psychology, 16,* 67–68.

Sciutto, M. J. (2000). Demonstration of factors affecting the *F* ratio. *Teaching of Psychology, 27,* 52–53.

MULTIPLE-CHOICE QUESTIONS

1. In an analysis of variance comparing two groups, the resulting *F* value is equal to
   a. t
   b. Cohen's *d*
   c. $\sqrt{t}$
   * d. t^2 (258)

2. The effect size of an *F* ratio is only interpretable if there is _____ *df* in the numerator.
   a. 0
   * b. 1
   c. 2
   d. more than 2 (259)

3. In the *F* test, one is analyzing _____ rather than analyzing _____.
   a. variances; error
   b. means; error
   c. means; variances
   * d. variances; means (264)

4. Variance is also known as
   a. the *F* ratio.
   b. the sum of squares.
   * c. the mean square.
   d. the standard deviation. (261)

5. The *F* ratio will usually have a value close to 1.0 when the variation between treatment conditions is _____ the variation within the treatment conditions.
   a. smaller than
   * b. similar to
   c. greater than
   d. eliminated by (259-260)

6. The total sum of squares is equal to the between-conditions sum of squares _____ the within-conditions sum of squares.
   * a. plus
   b. minus
   c. multiplied by
   d. divided by (261)

7. The _____ *SS* is defined as the sum of squares of deviations from the grand mean for all of the scores.
   * a. total
   b. between-conditions
   c. within-conditions
   d. grand (261)

8. The _____ *SS* is defined as the sum of squares of deviations of the condition means from the grand mean.
   a. conditional means
   b. total
   * c. between-conditions
   d. within-conditions (261)

9. The _____ *SS* is defined as the sum of squares of deviations of the individual measurements from their condition means.
   a. error
   * b. within-conditions
   c. total
   d. between-conditions (261)

10. For his experiment, John has three different conditions with five subjects per condition. What are the *df* between-conditions for John's study?
   a. 15
   * b. 2
   c. 14
   d. 12 (263)

11. In John's experiment, there are three different conditions with five subjects per condition. What are the degrees of freedom within-conditions for this experiment?
    a. 2
    b. 14
  * c. 12
    d. 15 (263)

12. John has three different conditions in his experiment, with five subjects in each condition. What is the total number *df* associated with John's study?
    a. 2
    b. 12
  * c. 14
    d. 15 (263)

13. To calculate the mean square, the sum of squares is divided by the _____ associated with it.
    a. *F* test
    b. *p* value
  * c. degrees of freedom
    d. error (263)

14. The *F* ratio is determined by dividing the _____ mean square by the _____ mean square.
    a. total conditions; between-conditions
    b. within-conditions; between-conditions
    c. within-conditions; total conditions
  * d. between-conditions; within-conditions (264)

15. When determining the probability of an obtained F, one must take into account both the *df* _____ and the *df* _____.
    a. total; within
    b. within; total
    c. between; total
  * d. between; within (264)

16. *F* is intrinsically a _____ test of significance.
    a. precise
  * b. one-tailed
    c. two-tailed
    d. preferable (264)

17. Computing a one-way ANOVA, Janet learns that there is a significant difference among the four groups in her study. To reveal precisely which group or groups differ from the rest, she must
    a. construct a BESD.
  * b. do tests of simple effects.
    c. calculate the effect sizes and corresponding confidence intervals.
    d. compute another ANOVA. (266)

18. A research design in which there are two levels for each of two factors is known as a(n) _____ design.
a. one-way
b. simple effects
* c. two-way
d. interaction (267)

19. The overall effect that an individual factor has on the dependent variable is also referred to as a(n)
a. simple effect.
* b. main effect.
c. interaction effect.
d. experimental effect. (268)

20. The residual or leftover effect after the main effects have been removed from the grand mean is known as the
a. simple effect.
b. error.
c. experimental effect.
* d. interaction effect. (268)

21. Error is considered to be the deviation of a score from the
a. grand mean.
* b. mean of the condition.
c. mean of the row.
d. mean of the column. (271)

22. Focused comparisons of more than two conditions are also known as
* a. contrasts.
b. omnibus *F* tests.
c. tests of simple effects.
d. lambda coefficients. (273-274)

23. When computing a contrast, the sum of the lambda coefficients must equal
* a. zero.
b. one.
c. the number of treatment conditions.
d. the *F* value. (274)

24. The correlation between the group means and their respective contrast weights is known as the
a. contrast correlation.
* b. alerting correlation.
c. effect size correlation.
d. coefficient of determination. (276)

25. Designs that require participants to be measured more than once in order to answer the research question are referred to as _____ research.
a. nonintrinsically repeated measures

b. omnibus statistical procedures
c. focused statistical procedures
* d. intrinsically repeated measures (277-278)

26. In order to increase the statistical power of her study, Judy decides that she will administer several treatments to each of her subjects. Judy is using _____ research in order to address her question of interest.
* a. nonintrinsically repeated measures
b. intrinsically repeated measures
c. simple effects
d. analysis of variance (281)

SHORT ESSAY QUESTIONS

1. How is the *F* test similar to the *t* test? Under what conditions can either test of significance be used?

2. Why is the *F* test considered to be a test that analyzes variances rather than a test that analyzes means?

3. How is the "signal-to-noise" analogy applicable to understanding the logic behind *F* tests?

4. What does the "sum of squares" represent? How is this related to the "mean square"?

5. What does the *p* value associated with an *F* ratio represent?

6. Why are effect sizes not calculated for *F* ratios that have more than 1 *df* in the numerator?

7. Why might a researcher sometimes prefer a two-way design to a one-way design?

8. What does error represent in the analysis of variance?

9. Why can individual scores be conceptualized in terms of an additive model?

10. What does the interaction effect in a two-way ANOVA represent?

11. What is the difference between effect size *r*s, contrast *r*s, and alerting *r*s?

12. What is the difference between intrinsically repeated measures research and nonintrinsically repeated measures research?

CHAPTER 15: *THE ANALYSIS OF FREQUENCY TABLES*

CHAPTER OUTLINE

I. What Is the Purpose of Chi-Square (χ^2)?

A. **Chi-square** (symbolized as χ^2) is a statistic, like *t* and *F*, that indicates how unlikely it is that the relationship investigated has occurred by chance. Like *t* and *F*, it does not immediately indicate the strength of the relationship between the variables.

B. Chi-square is computed for tables of independent frequencies (also called **counts**) and therefore can be thought of as a comparison of counts.

C. Chi-square tests the relation between two variables by assessing the discrepancy between the theoretically **expected frequency** (f_e) and the obtained or **observed frequency** (f_o).

D. Chi-square differs from *t* and *F* in that it can be used for dependent variables that are not scored or scaled.

E. Like *F*, chi-square can be a focused or an omnibus test. Chi-squares with 1 *df* are focused tests, whereas those with $df > 1$ are omnibus tests.

II. How Do I Compute 1-*df* Chi-Squares?

A. The general formula for chi-square is:

$$\chi^2 = \sum \frac{f_o - f_e}{f_e}$$

where

1. f_o is the observed frequency in each cell.
2. f_e is the expected frequency in that cell, which is determined by using the following equation:

$$f_e = \frac{(\text{Column total})(\text{Row total})}{\text{Grand total}}.$$

B. If the null hypothesis of no relation between the rows and columns were true, one would expect the f_o and f_e to be similar in magnitude. Observed frequencies that are substantially larger and smaller than the expected frequencies would cast doubt on the null hypothesis.

III. How Do I Obtain the *p* Value, Effect Size, and Confidence Interval?

A. Similar to *t* and *F*, there is a different chi-square curve for every value of the degrees of freedom.

B. The degrees of freedom (*df*) of chi-square are defined as

$$df = (\text{rows} - 1)(\text{columns} - 1)$$

C. As χ^2 becomes larger than the *df*, the null hypothesis becomes less likely to be true.

D. The effect size is computed using the **phi coefficient** (see Chapter 11).

E. Using the procedure described in Chapter 12, one can compute a 95% confidence interval around the observed effect.

IV. What Is the Relationship Between 1-*df* χ^2 and Phi?

A. If the sample size (N) is not too small ($N > 20$), and if the smallest expected frequency is not too small (e.g., less than 3 or so), one can test the significance of phi coefficients by chi-square tests, because

$$\chi^2 = (\phi^2)(N)$$

B. χ^2, (like t and F) is the product of the effect size and the study size. So, the larger the effect or the more sampling units in the chi-square table, the greater will be the value of the χ^2.

C. To obtain the value of the effect size correlation (phi) from the 1-*df* chi-square, use the following formula:

$$r_{\text{effect size}} = \phi = \sqrt{\frac{\chi^2}{N}}$$

V. How Do I Deal With Tables Larger Than 2 × 2?

A. When there are many cells in a chi-square table, a statistically significant chi-square may be more difficult to interpret than in a 2 × 2 table.

B. In these situations, statistically significant chi-square values only indicate that somewhere in the data the observed frequencies depart noticeably from the expected values, but not where the difference may be found.

C. One can closely examine the deviations associated with each cell that contributed the most to the overall large chi-square.

D. Using a procedure called **partitioning of tables**, one can subdivide the large table of counts into smaller (e.g., 2 × 2) tables. Additional chi-squares are then computed based on portions of the overall table.

1. Either a prior theory or hypothesis or the nature of the results can guide one's judgments as to which additional chi-squares to compute.
2. The number and size of the subtables are guided by certain statistical rules, and the calculations also require certain statistical adjustments.

E. One can use a procedure known as **standardizing the margins** (see below).

VI. How Is Standardizing the Margins Done, and What Can It Tell Me?

A. In the context of the procedure referred to as *standardizing the margins*, standardizing means that uniform row margins and uniform column margins are produced.

1. This procedure, through an iterative process, allows one to set all of the row totals equal to each other and all of the column totals equal to each other.
2. By taking the margins into account, one can more easily assess which cells are overrepresented and underrepresented in the frequency table.

VII. What Is a Binomial Effect-Size Display Used For?

A. The **binomial effect-size display** (**BESD**) is another approach that is predicated on the assumption of uniform marginal values.

B. The BESD can be used to provide an idea of the practical implications of an effect size indexed by a correlation coefficient.

1. It is called a display because it converts the "success rates" in the experimental and control groups into a 2 × 2 table.

2. It is called binomial because two variables are presented as dichotomous.

C. The BESD is obtained from any effect size *r* simply by computing the treatment success rate as 100(.50 + *r*/2) and the control condition success rate as 100(.50 – *r*/2).

C. Having the rows and columns always sum to 100 makes the values in the A, B, C, D cells easier to interpret and compare as proportions or percentages.

D. In the text's example, the effect size *r* was the phi coefficient, but the BESD can also be used with the point-biserial *r* computed from an independent-sample *t*, with the partial *r* computed from a paired *t*, and with the contrast *r*, alerting *r*, and effect size *r* associated with contrasts on three or more groups. However, the interpretation is more subtle than in a two-group design (for further discussion, see Rosenthal et al., 2000).

VII. A Journey Begun

A. Whether the conclusion of this chapter represents the start or the end of students' journey in behavioral or social (or some other area of) research, students should now have a deeper understanding of the applicability and limits of the scientific method.

LECTURE IDEAS AND ACTIVITIES

1. A variation on an exercise described by William Hunter (1981) can be used to introduce chi-square. Pose the question to students, "Does the government mint the same number of pennies and dimes each year?" Using as the null hypothesis that the same number of pennies and dimes were minted each year during a four-year period, have students examine the change in their pocket, counting the number of dimes and pennies they have that were minted during a specified four-year period. Based on the students' counts, a chi-square can be computed to test the hypothesis of no difference. Because the resulting chi-square will have more than one *df*, the instructor can then shift the discussion to procedures such as standardizing the margins that can be used to interpret larger tables of counts.

Hunter, W. J. (1981). Hypothesis testing—to "coin" a term. In L. T. Benjamin, Jr. and K. D. Lowman (Eds.), *Activities handbook for the teaching of psychology* (Vol. 1, pp. 16–17). Washington, DC: American Psychological Association.

MULTIPLE-CHOICE QUESTIONS

1. If one wanted to determine immediately whether an observed relation has occurred by chance or not, one could use all of the following statistics EXCEPT for
   a. F
   * b. r
   c. χ^2
   d. t (288)

2. Which of the following statistics is useful as a significance test for examining tables of counts?
   a. r
   b. t
   * c. χ^2
   d. F (288)

3. Stacy is interested in determining whether there is a relation between one's year in school (i.e., freshman, sophomore, junior, or senior) and whether one is participating in on-campus activities. Which statistical test should Stacy use?
   a. Independent *t* test
   * b. Chi-square
   c. *F* test
   d. Correlated *t* test (289)

4. Linda wants to know whether there are equal numbers of men and women in the different majors at her college. Which statistic should Linda use to answer this question?
   a. r
   * b. χ^2
   c. t
   d. F (289)

5. Chi-square assesses the discrepancy between the observed frequencies and the _____ frequencies in a table of counts.
   a. ideal
   b. exact
   * c. expected
   d. desired (289-290)

6. What is the expected frequency in the right-hand upper cell of the following table?

4	2
6	8

   a. 4
   * b. 3
   c. 7
   d. 2 (290)

7. The number of degrees of freedom associated with a chi-square based on a 2 × 3 table of counts is
   a. 1.
   * b. 2.
   c. 6.
   d. Unable to determine without knowing *N*. (291)

8. The obtained chi-square value must be _____ the degrees of freedom before one can begin to doubt the null hypothesis.
   a. smaller than
   b. equal to
   * c. larger than
   d. divided by (291)

9. Which effect size index is used with the 1 *df* chi-square?
   a. Fisher *z*
   b. point biserial *r*
   c. Cohen's *d*
   * d. phi coefficient (292)

10. John finds a significant chi-square value with 2 *df*. What should he do next?
   a. Calculate the effect size.
   b. Assess the effective power of the study.
   * c. Use a procedure for interpreting large tables of counts.
   d. Determine the exact *p*-value associated with the chi-square value. (288-294)

11. While trying to interpret a 3 × 4 table of counts that yielded a significant chi-square value, Mark examines how much each individual cell entry contributed to the overall chi-square value. Mark is looking for _____ values under the premise that such values would be unexpected.
   a. close to zero
   b. small
   c. equivalent
   * d. large (294)

12. A procedure for interpreting large tables of counts that involves computing additional chi-squares on portions of the overall table is referred to as
   * a. partitioning of tables.
   b. standardizing the margins.
   c. identifying unexpected cell values.
   d. partitioning of effect sizes. (295)

13. The procedure in which one tries to set all the row totals equal to each other and all the column totals equal to each other in order to interpret a statistically significant chi-square with more than 1 *df* is called
   a. partitioning of the tables.
   b. identification of unexpected cell frequencies.
   * c. standardizing the margins.

d. equalization of the cell frequencies. (295)

14. One useful tool for evaluating the practical importance of an observed result is the
a. statistical significance test.
b. statistical power analysis.
* c. binomial effect size display.
d. Type I vs. Type II ratio. (296)

15. The binomial effect size display can be used to evaluate the _____ an observed result.
* a. practical importance of
b. statistical significance of
c. statistical importance of
d. effect size associated with (296-298)

SHORT ESSAY QUESTIONS

1. Under what conditions would one use a chi-square rather than a *t* or *F* test to test the null hypothesis?

2. How does chi-square test the relation between two variables in a table of counts?

3. How are the chi-square and the phi coefficient related?

4. What does a significant chi-square with more than 1 *df* indicate? What does it not indicate?

5. What procedures could one use to interpret larger tables of counts that produced statistically significant chi-square values? How do these procedures help in the interpretation of the table of counts?

6. What is the purpose of the binomial effect size display (BESD)? How does one construct and interpret a BESD?